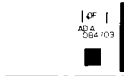
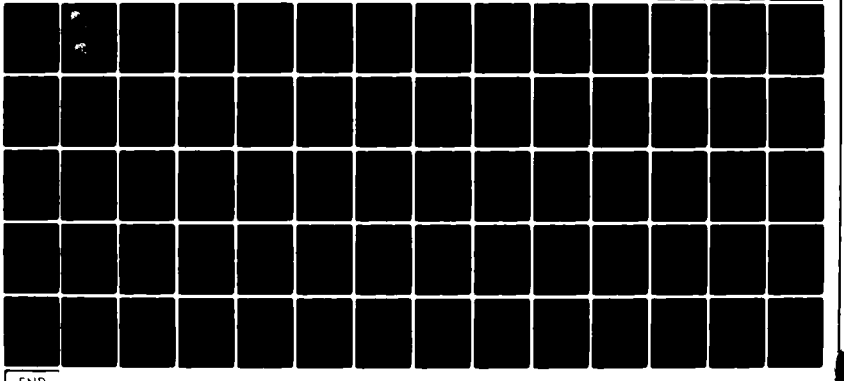
AD-A084 703 SPEECH COMMUNICATIONS RESEARCH LAB LOS ANGELES CA F/G 17/2
CRITICAL ISSUES IN AIRBORNE APPLICATIONS OF SPEECH RECOGNITION.(U)
1979 W A LEA N62269-78-M-3770

UNCLASSIFIED NL





END
DATE FILMED
6-80
DTIC

CRITICAL ISSUES IN AIRBORNE APPLICATIONS OF SPEECH RECOGNITION

Wayne A. Lea

Speech Communications Research Laboratory
806 West Adams Boulevard
Los Angeles, CA. 90007

FINAL TASK REPORT

NAVAL AIR DEVELOPMENT CENTER

CONTRACT NUMBER N62269-78-M-3770

Submitted to

LCDR Norman E. Lane (Code 6041)
Naval Air Development Center
Warminster, PA. 18974

The views and conclusions expressed in this document are those of the author alone (or of the experts he has polled, whenever so indicated), and should not be interpreted as necessarily representing the official policies of the sponsor or the U.S. Government.

CRITICAL ISSUES IN AIRBORNE APPLICATIONS OF SPEECH RECOGNITION

Wayne A. Lea

Speech Communications Research Laboratory
806 West Adams Boulevard
Los Angeles, CA. 90007

FINAL TASK REPORT

NAVAL AIR DEVELOPMENT CENTER

CONTRACT NUMBER N62269-78-M-3770

Submitted to

LCDR Norman E. Lane (Code 6041)
Naval Air Development Center
Warminster, PA. 18974

The views and conclusions expressed in this document are those of the author alone (or of the experts he has polled, whenever so indicated), and should not be interpreted as necessarily representing the official policies of the sponsor or the U.S. Government.

EXECUTIVE SUMMARY

The purpose of this study was to identify critical issues and problems in the airborne use of speech recognizers (i.e., the use of spoken commands to control airborne systems), and to help guide the continued development of an adequate speech recognition technology for the 1980 to 1985 time frame. This study grew out of a need to apply the results of a recent comprehensive survey of speech recognition technology (Lea and Shoup, 1979) to the specialized conditions of airborne crewstation tasks. The general recommendations of that previous review have been specialized and expanded upon in this study project, and further "at-the-desk" studies were done of the airborne conditions, the needed technology and the priorities of specific problems, and the practical ways of resolving remaining issues through a coordinated program of future projects.

The earlier general survey by Lea and Shoup reiterated the now-fairly-well-known advantages and disadvantages of voice input of commands. Speech is recognized as a fast, natural additional modality which frees the human's hands and eyes for other tasks and requires little panel space or complex apparatus. Yet, spoken versions of the same intended message will vary significantly from time to time and from talker to talker, and speech is subject to interference from noise, distortions, and special airborne conditions such as vibration, g forces, the oxygen mask, stress and fatigue, and size and weight limitations. Also, there are questions of user acceptance of this new technology, and the criticality associated with making errors in spoken command and control of airborne systems. There is presumably a limited "market" for airborne speech recognizers, so that substantial development costs may have to be borne by a moderately small number of operational units. This author contends that the problems with voice input in airborne situations can be overcome, and that a vigorous effort in airborne applications of speech recognizers is warranted. In operational airborne situations where the crew member's usual manual and visual communication channels are overloaded, the naturalness, freedom of movement, and efficiency of the voice modality become particularly significant. The airborne voice input task is limited on some important dimensions such as the limited speaker population that must be handled, the highly trained character of the airborne users, and the clearly defined limited tasks of crewstation activity.

The 27 year history of work on speech recognition has addressed a broad spectrum of types of recognizers, ranging from very accurate isolated word

recognizers, up through limited forms of connected speech recognition such as digit string recognizers and recognizers of formatted word sequences, and extending up to the ambitious research efforts on systems that can understand total sentences of spoken English. Current commercial devices show that speaker-dependent isolated word recognition is a practical, accurate technology that is having substantial commercial success. Digit string recognition is now becoming commercially available, and limited forms of formatted word-sequence recognition is not far behind. These types of recognizers will clearly impact the operational airborne applications in the next few years. More advanced and versatile forms of continuous speech recognition will require further work, and should begin to be operationally applied near the end of the 1980-1985 time frame. The ARPA Speech Understanding Research project demonstrated limited successes in machine understanding of spoken sentences, and showed the potential for further advancing the versatility of speech recognizers.

Significant "gaps" still remain in speech recognition technology, and these unresolved problems or limitations of current technology need to be addressed by further research and development work, as well as specific projects in advanced system development, evaluation and refinement of current devices, and transfer of technology to practical applications. There is currently a substantial gap between practical isolated word recognition and the demonstration versions of research systems that handle continuous speech. Also, for successful versatile spoken interaction with machines, crucial work is needed on the acoustic phonetic, prosodic, and phonological analyses that form the "front end" of powerful recognizers. Procedures are needed for the total evaluation of recognizer system performance, including accuracy statistics, origins of errors, assessment of task complexity, etc.

The definition of these and other "gaps" prompted Lea and Shoup to define a list of needed project types, which has influenced the recommendations in this report. However, specific work had to be done to consider the implications of <u>airborne conditions</u> on recognition work. Previous studies have shown that recognizers do suffer from the introduction of noise, vibration, g forces, slipping of the oxygen mask, stress and fatigue in critical mission phases, and the other demands of (simulated) airborne situations. Boeing/Logicon assessed the various crewstation tasks aboard a P-3C antisubmarine aircraft, and defined an overall assessment of the technical feasibility, usefulness, criticality, and other variables involved in use of voice for each crewstation task. From their assessment came definitions of tasks that should get early

priority in the introduction of voice technology, and they recommended 25 projects to advance the airborne and training applications of speech recognizers. Other work has been and is being conducted on airborne applications, by NADC, NASA Ames, RADC, and by the Royal Aircraft Establishment in Britain plus other groups in France and other NATO countries. Experiments have repeatedly shown the utility of voice, in contrast to other modalities, for complex data entry tasks and situations where the user is simultaneously occupied with other tasks. Several illustrative prototype isolated word recognition facilities have been developed for flight applications, at NADC, NASA Ames, and elsewhere. NASA Ames studies showed the significance of vocabulary size, and NADC studies have shown effective procedures for using syntax, semantics, and logical command associations to aid recognition.

All this previous work has defined a variety of remaining issues, which can be organized into major categories according to stages in the process of communicating messages to machines. Those general issue categories include: task conditions and the selection of the most appropriate tasks; human factors, such as the value of speech and the actual need for continuous speech; language issues, such as vocabulary size and confusability, the syntax of command sequences, etc.; environmental conditions and channel characteristics, such as noise, distortions, g-forces, etc.; recognition techniques, such as front-end processes of acoustic, phonetic, prosodic, and phonological analysis, and further research and development work in recognizer designs; performance evaluation, such as accuracy, identifying sources of errors, comparative evaluations of recognizers, and enhanceabilities of techniques; and response procedures, including methods for error detection and correction. The reader is encouraged to study Figures 8 (page 41) and 9 (page 54) at this time, to obtain a summary of these issues and this author's assessments of their priorities.

In general, among the most critical issues are those of task selection, evaluation of the best types of speech to use, assessment of vocabulary confusability, analysis of noise effects on recognition, and research and development work on advanced techniques in acoustic phonetic analysis (that is, the detection of the vowels and consonants that are in the spoken utterance). Close behind such high priority issues are others dealing with: stress and fatigue effects; the evaluation of the actual need for continuous speech; research and development work on prosodics (intonation, rhythm, and stress patterns that show the structures of sentences); phonological analysis

procedures (that account for the slurred form of smooth-flowing speech); normalization adjustments for different channels, speakers, and rates of speaking; and the development of databases, testing conditions, and procedures for comparative evaluations of speech recognizers. All the issues in Figure 9 (page 54) deserve consideration.

The need to resolve these critical issues has lead this author to suggest the program plan of 22 projects as summarized in Figure 10 (page 56 of this report). Projects of high priority are suggested on: the selection and evaluation of airborne tasks; the analysis of airborne environmental effects such as noise and stress and fatigue; the evaluation of the most suitable types of speech for airborne tasks; the definition of standard databases, benchmark tasks, and comparative evaluation procedures, for assessing recognizers; development of a methodology for assessing confusabilities of vocabularies; and long-range research on acoustic phonetic, prosodic, and phonological analysis aspects of recognizers, coupled with time, channel, and speaker normalization procedures. Moderate priority is given to: other airborne environmental conditions; later refinements and applicability studies for recognizers; development of a spectrum of moderately challenging to boldly ambitious continuous speech recognizers; sophisticated language, task, and system evaluation and enhancement procedures; and higher-level linguistic aspects of versatile recognizers. Other lower priority efforts are also suggested. This ambitious program is estimated to cost about $2 million per year over the 1980-1984 time frame, and might be undertaken by the cooperative efforts of several government agencies and contractors. The program appears to appropriately apply current and projected states of technology to the operational needs in airborne applications. The potential value of voice communication with airborne systems warrants immediate attention to such a coordinated program for transferring on-going research and development work into practical airborne applications.

TABLE OF CONTENTS

1. INTRODUCTION

Machine recognition of speech has apparently "come of age", as evidenced for example by the title of a recent article in Science magazine, which asserted that "More people are talking to machines, as speech recognition enters the real world" (Robinson, 1979). Speech recognition devices have begun to impact the commercial world, with considerable user satisfaction. The field shows some signs of rapid expansion, partly resulting from an expanded spectrum of available devices and an expectation of substantially lower costs in the next few years. One applications consultant projects a ten-year market of $1.5 billion or more (Nye, 1979). Despite such promising commercial trends, almost all previous military work has been confined to preliminary experiments with prototype systems. Some valuable studies have been done with various military applications (e.g., cartography, recognition and voice validation for access to secure areas, and training air traffic controllers and other skilled communicators) and with realistic operational conditions (noisy low-fidelity channels, and simulated airborne flight with "g forces", vibration, oxygen masks, fatigue, etc.).

Airborne applications of speech recognizers are worthy of the considerable attention recently focused on them (Curran, 1978; Coler, et al., 1978; Feuge and Geer, 1978; Bridle and Peckham, 1978; Haton, 1979). This report attempts to relate such specific applications work to the general context of research and development work and commercial efforts in speech recognition, and explores the many issues that must be addressed to assure a true contribution of speech recognizers to mission effectiveness. The remainder of section 1 presents the purposes, background, and methodology of this brief project conducted by Speech Communications Research Laboratory. Section 2 reviews the general historical context and current problem areas in speech recognition, section 3 discusses some relevant aspects of cockpit applications, section 4 outlines the issues pertinent to effective airborne uses of recognizers, and section 5 summarizes recommendations regarding future work that should be pursued to assure effective airborne uses of speech recognizers.

1.1 The Goals of this Study

The purpose of this study was to identify critical issues and specific problems that are likely to be encountered when automatic speech recognition is employed in operational aircraft crewstations. This review of issues and problems will hopefully serve as a precursor and guide to continued advanced development of airborne voice input facilities. Attention is given to the implications of projecting current technology into the 1980-1985 time frame. For each identified issue or problem, the potential severity and the likelihood of resolution will be discussed, based on

current evidence bearing on that specific problem or issue. Recommendations are given for further work that is needed, along with some estimates of the amount of effort required for each topic.

It should be noted that this study was conducted by a researcher who has been active in government, industrial, and laboratory projects for over a dozen years, but who admittedly is not an expert on airborne conditions, military requirements, or details of manufacturing or marketing commercial recognizers. It was well beyond the scope of this study to investigate, through simulations or operational cockpit situations, the various airborne crewstation tasks and their needs for voice commands. Recent work (Feuge and Geer, 1978), reviewed briefly in section 3 of this report, has investigated the crewstation tasks aboard a P-3C aircraft, and has suggested 25 potential projects and their priorities. Other earlier research (Martin and Grunza, 1974; Montaque, 1977; Curran, 1978) has investigated the effectiveness of speech recognition devices under such airborne conditions as wearing an oxygen mask, high noise levels, and undergoing acceleration ("g-forces").

The current study primarily adds assessments based on the author's knowledge of current and projected capabilities in speech recognition.

1.2 Background of this Study

Following the completion of the 5 year, $15 million ARPA SUR project (to be described in section 2.2 of this report), Dr. June Shoup and this author (Wayne A. Lea) were contracted to review the contributions of that project, and to survey the total current technology in speech recognition. Following a comprehensive survey, including conferring with over 100 experts in all aspects of speech recognition technology, Lea and Shoup prepared a final report which provided general recommendations for further work to be done in all aspects of recognition, ranging from specific applications to needed research. This is apparently the most recent and general survey of speech recognition technology, and the reader is encouraged to read the final report of that project (Lea and Shoup, 1979). One major recommendation discussed when the review was presented at an ONR Workshop on Speech Recognition in June, 1978, was the need to relate the general review results from that study to specific military operational settings. The study described in <u>this</u> report is an initial attempt to specialize and expand upon the earlier general conclusions, for operational airborne applications. Though the reader can gain extensive general insight from the earlier Lea and Shoup report, <u>this</u> report stands alone as an independent document concerning the critical issues in airborne use of speech recognition technology.

1.3 Methodology

The methodology used during this study is not unusual for "at the desk" studies of practical problems. It consisted of: (1) selecting from the very general conclusions of the earlier survey those aspects which directly apply to aircraft; (2) studying available documentation concerning relevant conditions aboard representative aircraft; (3) conferring with selected military representatives interested in aircraft applications of speech recognizers, and (4) relating the known operational requirements to the author's knowledge of advanced technology and future trends in both laboratory and practical recognition facilities.

With guidance from LCDR Norman Lane of the Naval Air Development Center, this study assumed the following guidelines for assessing future technology and operational requirements:

(1) recognition capabilities are projected into the 1980-1985 time frame;

(2) attention is focused on systems with 100 to 300 word vocabularies, with both isolated word recognition and limited continuous recognition (digit strings or carefully formatted word sequences);

(3) potential issues or problems to assess include (but are not restricted to): vocabulary size; discriminability of voice commands; multiple versus single speakers; extent of required training; environmental noise; system packaging (weight and volume); capability for real-time operation; effects of g loading, fatigue, and stress; equipment reliability; potential conflicts with ICS and radio traffic; microphone requirements and limitations; and command syntax and syntactical processing requirements;

(4) for each identified issue or problem, its potential severity, and the likelihood of resolution, will be indicated and related to the evidence bearing on that problem; and

(5) estimates are attempted concerning the nature of additional work required for resolution of each problem, and the magnitude of required effort.

Given these guidelines and a very limited level of effort, the author attempted to relate general knowledge of the known speech recognition technology to operational airborne applications. The reader is cautioned to keep in mind the limitations and lack of operational airborne experience reflected in this methodology.

2. GENERAL SURVEY OF SPEECH RECOGNITION TECHNOLOGY

This section summarizes the major technical accomplishments of previous work in speech recognition, including a large recent project which was conducted to demonstrate the technical feasibility of naturally speaking sentence-like commands to a computer, and having the machine "understand" the content of the spoken sentence sufficiently to produce a correct response. The purposes, history, and current capability in voice input to machines are briefly presented in section 2.2, to provide the context in which accomplishments can be evaluated, and to provide the background for the author's conclusions about future work that still needs to be done. Based on the general assessment of current technology as summarized in section 2.3, the gaps in current technology listed in section 2.4, and the broad scope of recommendations for further work in section 2.5, there is definite justification for exploring the application of this advancing technology to airborne applications, as will be discussed in subsequent sections of the report.

First, let us clarify a few terms used in this review. "Speech recognition", and the more specific concept of "speech understanding", as used in this report, refer to the process of a machine determining the content or message being conveyed by a human's spoken utterance. Speech recognition generally may refer to machine analysis and decision making based on the human user speaking perhaps only a single spoken word, or a short sequence of words (with very strict rules about which word can follow another word), or a full-blown natural-language sentence. Speech understanding, as used in this section, refers specifically to a machine's ability to receive total sentences, and recognize not only the wording of the sentence but also determine enough about the structure of the sentence and its "meaning", so that it can produce a correct response, such as answering an inquiry for data, or operating a mechanism, or performing a calculation.

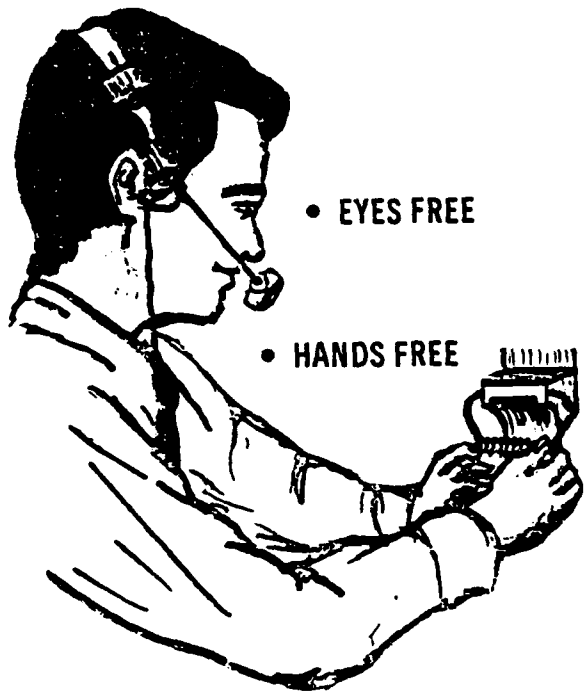
Speech recognition (that is, determining "what was said") may be distinguished from speaker recognition, wherein the machine determines who said the utterance. We do not consider speaker recognition, or other technical topics such as machine detection of the language being spoken (language identification), or machine analysis of how or where the utterance was spoken (i.e., machine detection of the talker's emotional stress, the talker's dialect, or the noisy environment in which he or she spoke). Similarly, we do not consider speech output from the computer (so-called speech synthesis or voice response), except as how it influences methods of speech input. The goal of speech recognition is to understand enough of the linguistic message being spoken by the human, so as to provide a correct and useful machine response. The airborne situation, and the limitations of current technology, will

place certain constraints on recognition systems, as we shall see in later sections.

2.1 Advantages and Disadvantages of Voice Control of Machines

One reason for growing interest in speech recognition is the growing importance of machines (especially the modern digital computers) in our lives. A critical problem in the effective use of machines concerns the ease and accuracy with which commands are communicated. This need for the human user to effectively instruct the machine is clearly evident in large control consoles that are used with navigation systems for ships and aircraft and with other complex command and control systems. However, effective communication of commands is also needed for the best use of small data-entry devices with only a few knobs or switches. Errors in pushing the wrong button, dialing or tuning wrong, or using the wrong sequence of commands can be wasteful of time and effort, frustrating to personnel, and sometimes directly hazardous. In addition to being prone to errors, the unnatural ways of communicating can be tedious and expensive (in that they require highly trained personnel). For years, there has been a growing interest in versatile, "natural-to-the-human" techniques for human communication to computers and other machines. Computer specialists have been developing a variety of devices, programs, and "natural" programming languages for rapid interaction between humans and machines. Conversing with a machine in spoken form is one of the more challenging, but clearly most natural, means of instructing a machine.

Figure 1(a) summarizes some advantages of voice control of machines, and Figure 1(b) lists some disadvantages. Speech input to machines offers an unprecedented mobility to computer users, in that the user need not be in actual physical contact with the machine, but rather may walk around, turn aside, and, most importantly, use his hands and eyes for other tasks while speaking instructions to the machine. Speech is fast, natural, and nearly universal among humans, in contrast to the cumbersome, unnatural, and technically complex input techniques that have typically been used in commanding machines. Indeed, experiments with humans interacting to cooperatively solve various problems have shown that people can find solutions twice as fast when they are allowed to converse in spoken form, in contrast to when they communicate by typewriter, handwriting, or visual signs (Chapanis, et al., 1977; Ochsman and Chapanis, 1973). What makes the development of speech recognizers so hard is that different people speak in different ways, the noisy environment may interfere with reliable interpretation of the acoustic speech signal, and even the same single talker will vary from time to time in his pronunciations. Currently, speech recognition is a relatively costly way to communicate to machines, but that is expected to change soon; also, higher initial cost is often compensated



- HUMAN'S MOST NATURAL MODALITY
- LITTLE OR NO USER TRAINING
- PERMITS FAST, MULTIMODAL COMMUNICATION
- PERMITS SIMULTANEOUS COMMUNICATION WITH MACHINE AND OTHER HUMANS
- FREEDOM OF MOVEMENT AND ORIENTATION
- NO PANEL SPACE OR COMPLEX APPARATUS
- COMPATIBLE WITH TELEPHONE AND RADIO

(a) Advantages of voice input to machines.

- VOCABULARY (SIZE, CONFUSABILITY)
- TRANSDUCER AND CHANNEL CHARACTERISTICS
- SPEAKER VARIABILITY SEX, DIALECT, EXPERIENCE
- NOT PRIVATE
- ENVIRONMENTAL NOISE AND DISTORTIONS
- CURRENTLY COSTLY AND RESTRICTED

(b) Problems with voice input.

Figure 1. Advantages and disadvantages of voice input to machines.

for by reduced operating costs and personnel expenses resulting from more effective work load sharing with machines.

The idea of talking to machines is not exactly new. For about three decades, work has been done on some of the simplest ways of determining machine responses by speaking simple commands. Necessarily, almost all of that work was done on the highly restricted form of spoken input known as "isolated words". Single words (or short phrases that are treated as single words) are spoken in isolation, with easily detected pauses (silences) before and after each word. The reason for using isolated words is perhaps obvious. The pauses act as boundary markers, clearly showing where a word begins and ends. Then the pattern of speech wave between those time markers can be analyzed. This analysis can be done without having to consider the effects of surrounding words such as would be required in naturally flowing sentences, wherein each word gets slurred somewhat and distorted by the need of the human's tongue, lips, and other articulators to get ready for the next vowel or consonant to be spoken, and to gradually move away from the positioning needed for the previous sound. Thus, machine recognition of isolated spoken words is a lot like the recognition of single typewritten (or, more precisely, handwritten) words which are separated from neighboring words by clear spaces. Recognition of continuous speech is more complex in analogous ways to the relative difficulty of analyzing continuously flowing handwriting as contrasted to clearly-separated typewritten letters. In fact, recognition of continuous speech would be analogous to handwriting in which no spaces are left, even between words, but the lines flow directly from the end of one word into the beginning of the next word. It is obviously a major simplification to confine spoken inputs to single isolated words.

This analogy to handwriting analysis can be extended to illustrate other points about speech recognition. Just as different individuals have different handwriting, one can expect different speakers to have different "voice signatures", so that the same sentence or linguistic message may be expressed in different ways by different talkers. Machines may be designed to handle only one speaker's voice, or to readily adapt to many voices. Thus, we may distinguish between "speaker-dependent" and "speaker-independent" systems. Most work on isolated word recognizers has been confined to speaker-dependent systems that can handle one or a few speakers. Each new talker must then "train" the system to recognize his particular way of saying the alternative words. Also, just as an individual's writing may vary somewhat with fatigue, emotional states, style, speed of writing, and interference from a vibrating table, so speech is influenced by a tired vocal system, emotion, conversational versus formal style, rate of utterance, and environmental noise. Most speech recognizers have confined their applications to carefully spoken isolated

words in fairly quiet rooms, thus easing their recognition task as much as possible.

2.2 History of Speech Recognition

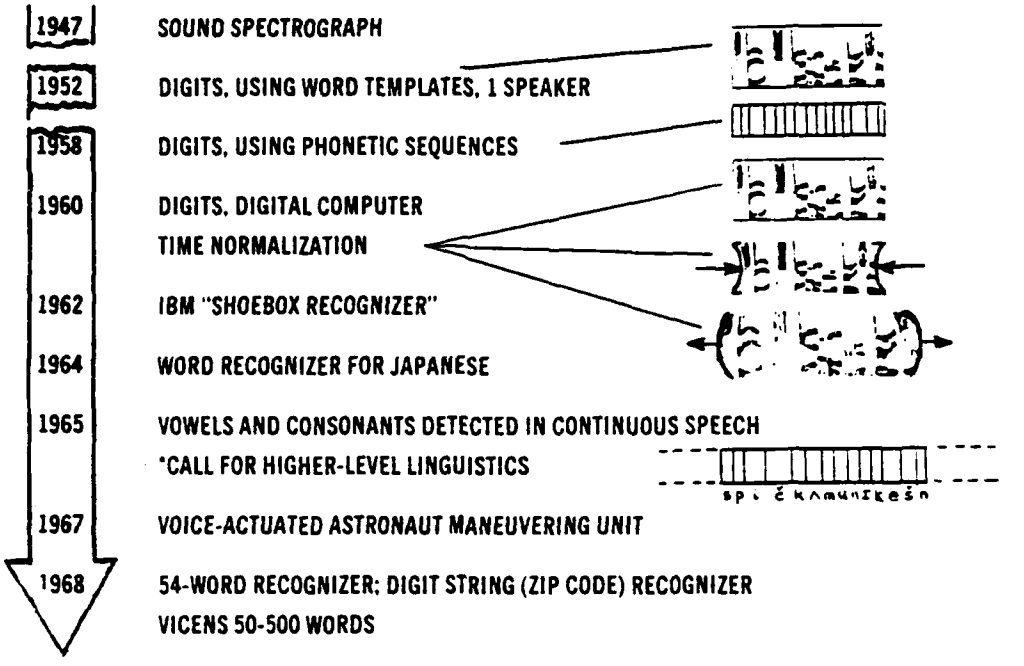
Reviewing the history of speech recognition work can help clarify the current capabilities and the conditions that have been producing improved abilities in vocal command of machines. Here we shall briefly consider work on recognizers of isolated words (section 2.2.1) and continuous speech (2.2.2), the large ARPA SUR project (2.2.3), and the resulting total spectrum of alternative types of recognizers (2.2.4). A pictorial summary of some of the highlights in the history of speech recognition is shown in Figure 2.

2.2.1 Work on Isolated Word Recognizers - Early word recognizers worked fairly well with simple comparisons or correlations of the total pattern for an incoming word with stored exemplars or templates. For example, in 1952, the first laboratory model of an automatic recognizer was developed (Davis, Biddulph, and Balashek, 1952), which identified which of the ten digits, zero to nine, was spoken by one specific speaker. That system used a two-dimensional template somewhat like the sound spectrogram shown in Figure 2. The two-dimensional array of numbers for an input word was cross correlated with stored training templates for each of the ten words, to decide which one of the ten words it was most similar to. Several years later (Dudley and Balashek, 1958), a method was used to segment the speech into phonetic units, or time slices, like vowels and consonants, and slightly better recognition scores were reported, for several talkers.

This early work was done with special purpose electronic hardware. The first work using a digital computer came in around 1960 (Denes and Mathews, 1960; Forgie and Forgie, 1959), along with the introduction of an important concept of time normalization, whereby short versions of an utterance that were spoken more rapidly than the training data were automatically stretched out or "normalized" to equal the normal duration of the training utterances, and slowly-spoken long versions could get reduced to a normalized length before comparisons and matching were attempted.

Throughout the 1960's, continued work was done on both the mathematical and the phonemic approaches to recognition, and initial attempts were made to recognize "continuous speech", such as word sequences without pauses between words. Important advancements were made in the classification of vowels and consonants occuring in continuous speech (Hughes, 1961; Hemdal and Hughes, 1965; Reddy, 1967). Also, major strides were made in detailed mathematical procedures for signal analysis, such as (a) the "fast Fourier transform" (which permitted rapid determin-

EARLY HISTORY OF MACHINE RECOGNITION OF SPEECH



RECENT HISTORY

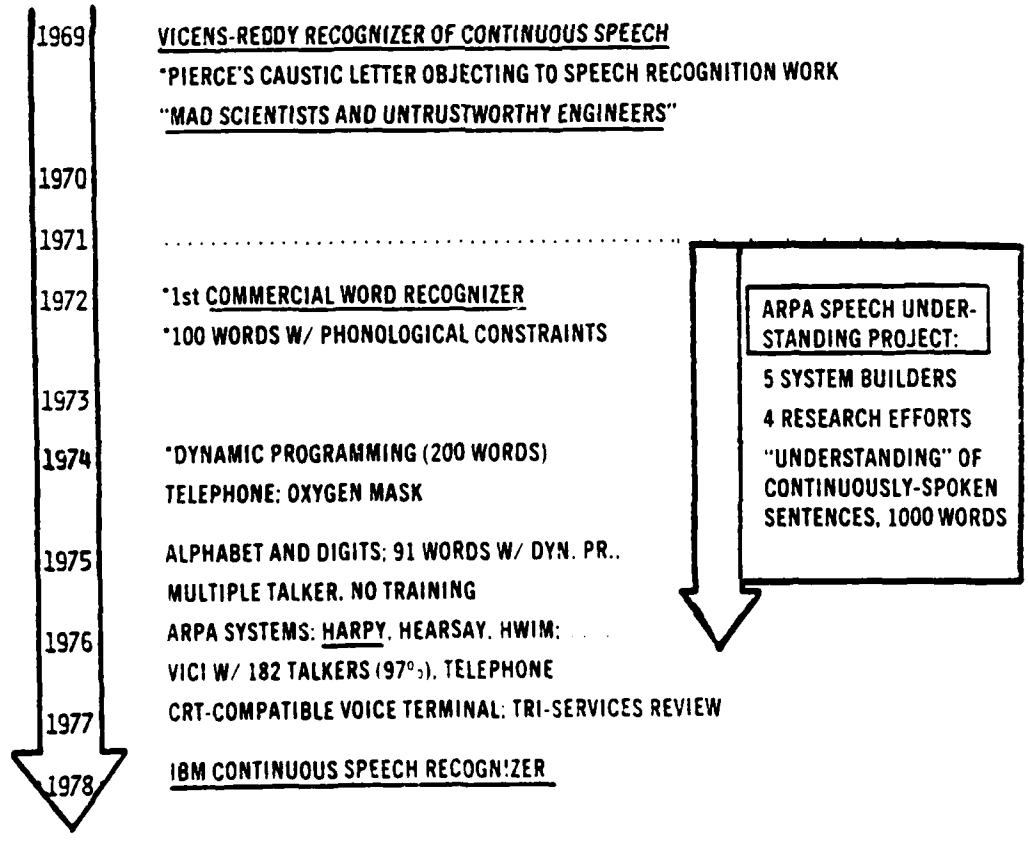


Figure 2. Some highlights in the history of speech recognition.

ation of the frequency spectrum in each short region of the speech, using a digital computer), and (b) "linear predictive analysis" (which separated the effects of the talker's vocal cords and his mouth-throat acoustic tube, so that formants and other interesting features of the speech signal could be more readily and accurately detected).

Of particular importance to the current study were some applications-oriented studies of isolated word recognizers in the presence of noise, g-forces, oxygen masks, and large speaker populations. Those studies will be summarized in section 3.

In 1972, the first commercial devices for isolated word recognition were offered, and since that time over 300 million words have been processed through such devices. About 400 practical devices have been sold (in addition to over 800 hobbiest devices recently marketed), and some of those systems are now providing 8, 16, or 24 hour service each day in factories, other commercial applications, and experimental military systems. The current status of commercial devices will be described in section 2.3.

Most commercial products work with small vocabularies of less than 30 alternative words or phrases, but some recognizers use syntactic constraints on allowable sequences of isolated words, to increase the effective vocabularies handled up to several hundred words. Thus, words from one small sub-vocabulary can appear first, then a "menu" of next acceptable words is possible, and so forth, yielding a discourse of human-to-machine interaction which can involve any of several hundred command words.

Laboratory studies with isolated word recognition schemes have also continued in recent years. In 1969 and repeatedly in 1972 and later years, Medress and his colleagues at Sperry Univac developed large vocabulary word or phrase recognizers using knowledge of English sound structure or phonology. Bell Laboratories researchers developed a series of recognizers, including a system for word recognition using dynamic programming and a spectral distance measure for comparing input patterns with stored word templates (Itakura, 1974; 1975). Itakura's metric and the dynamic programming algorithm have influenced many subsequent projects, including a study at Xerox (White, 1975) and the development of a commercial word-sequence recognizer at Nippon Electric Company (Tsuruta, 1978). More recently, other Bell Laboratories researchers have developed systems for handling: isolated and connected digits (Sambur and Rabiner, 1975, 1976; Rabiner and Sambur, 1975, 1976); travel reservations, using isolated words spoken over telephone lines (Schmidt and Rosenberg 1977); spelled speech and directory assisstance over the telephone (Lesk and McGonegal, 1976; Rosenberg and Schmidt, 1977); and other applications

(Levinson, et al., 1977; Sondhi and Levinson, 1977, 1978; Flanagan, et al., 1979). Bell Laboratories workers now claim that they could handle any voice of any American dialect, for limited isolated word recognition tasks, over the telephone (Rosenberg, et al., 1978), using only a few "templates" of alternative ways of pronouncing the words. Commercial firms are now actively considering speaker-independent recognition over telephone lines, and this is one of the primary goals of Dialog Systems, Incorporated.

Many other studies have been done with isolated words, and their technology may be considered to have "come of age", although costs are still high. We will discuss the current status of such limited recognizers more in section 2.3.

2.2.2 Continuous Speech Recognition- Early work in continuous speech recognition dates from as far back as the early 1960's, when attempts were made to recognize vowels and consonants as basic building blocks of continuous speech (Forgie and Forgie, 1959, 1962; Hughes, 1961; Hemdal and Hughes , 1965; Otten, 1965; Reddy, 1967).

The 1960's also saw the growth of an almost-universal call for the use of "higher level linguistic analysis" in speech recognizers (especially continuous speech recognizers), so that expected and grammatically-acceptable sequences of words would be used to limit the possible words that might be guessed to occur at various points throughout the utterance. Also, "semantic" constraints on meaningful sequences of words could be used to rule out some word sequences that might be hypothesized for the speech. One might also use known regularities of English sound structure or "phonology" to select the most likely words being pronounced. The basic contention was that as one moves from simple isolated words to the handling of continuously spoken sentences, all kinds of confusions in possible wording might arise, and the linguistic knowledge would help keep the alternative word sequences to be tried down to a minimum. In 1969, Reddy and Vicens introduced a recognizer that handled continuous speech, with a small vocabulary of 16 words, and a highly constrained syntax. The incorporation of linguistic knowledge seemed to be a formidable task, and at the end of that decade some influential researchers were pessimistic about the foreseeable future producing any adequate recognizers of continuous speech. For example, in 1969, the most popular letter to the editor that was ever published in the Journal of the Acoustical Society of America appeared. John Pierce (1969) of Bell Laboratories strongly objected to work in speech recognition and mused about its domination by "mad scientists and untrustworthy engineers." One might have expected that such caustic remarks by such an influential researcher would signal the end or a setback to speech recognition

work, especially since Pierce had earlier been instrumental in squelching the field of mechanical translation of languages. In some ways the field did flounder, but within several years Bell Laboratories itself was again doing its own work in speech recognition.

Other major advances were being made, however, in computer technology and "artificial intelligence" (the ability to perform tasks on machines that would be said to involve "intelligence" if humans did them). Procedures had been developed to have computers play games like checkers and chess, to make logical deductions and inferences, to recognize patterns such as handwritten characters or tanks in camouflage, and to rapidly search among thousands of alternatives to find the best solution to a problem. In addition, what have sometimes been called "friendly systems" were developed, which permitted users to very readily and naturally interact with a computer, without the need for cumbersome mathematical languages or unnatural diversions into the intricate details of inner workings of the machine. "Time sharing systems" permitted more than one user to effectively use a machine at the same time, and opened up the possibility for large groups of researchers to cooperatively work on various aspects of a complex problem (such as spoken sentence recognition), using the same computer system. The stage had clearly been set for major advances on the complex problem of continuous speech recognition.

2.2.3 The ARPA SUR Project - The Advanced Research Projects Agency (ARPA) apparently recognized this coalescing of all the essential ingredients for an effective assault on the task of understanding spoken sentences. ARPA had been influential in funding much of the relevant advances in artificial intelligence and advanced computer technology, and apparently could see the prospects for applying that work, and other recent advances in speech sciences and linguistic processing, to the recognition of speech. There apparently also was a keen awareness of the value of using speech understanding as a task which involved many sources of incomplete knowledge, each working together to help refine decisions about the content of an utterance. Systems could exploit recent advances in "syntactic parsing" (the determination of the grammatical structure or phrasal groupings in a sentence), "semantic analysis" (the interpretation of the "meaning" of a sentence), and "pragmatic analysis" (the determination of the appropriateness of a particular sentence in the context of previous discourse, in accord with the constraints of the task being performed by the human-plus-machine interactive system). The emerging theories of the sound structure of English ("phonological rules") and the initial attempts at characterizing the intonation, timing, rhythm, and accentual patterns of speech (so-called "prosodic structures") could be incorporated into the system. And all this could be coupled with

the advancing techniques in acoustic analysis and vowel and consonant identification. Alternative "control structures" existed for integrating all these processes into one cohesive system that carefully focussed on the most promising information and properly scheduled and coordinated all the subprocesses.

Thus, when ARPA commissioned a study group to explore the design of a large project for determining the feasibility of systems that understand speech, the conditions seemed ripe for integrating many disciplines into one cohesive effort. A dominant force in that study group was a collection of artificial intelligence experts who had been effective in previous ARPA projects dealing with other artificial intelligence tasks. This group defined an ambitious five year "Speech Understanding Research" (SUR) project, involving five initial contractors who were to build speech understanding systems. At the mid-term of the project, each contractor's intermediate test system was to be evaluated, and the best three or four systems were to be continued in development. A comprehensive set of system specifications were defined for the final evaluation of the resulting systems, as shown in Figure 3. These specifications did not necessarily represent "the last word" in necessary system performance conditions, but they were the best estimate of the study group concerning reasonable goals that would show the feasibility of speech understanding and would signal the emergence of a promising overall technology for the comprehension of continuously spoken sentences by complex systems. Despite the fact that at least one final system met or exceeded these goals, I believe that undue attention has been given to these system specifications. The primary initial goal (and the longest-lasting legacy) of this project was the successful demonstration of an emerging interdisciplinary technology for effective machine comprehension of continuously spoken sentences. Many problems in speech understanding were uncovered, and some promising initial solutions developed, but much more work is still to be done. To understand this, we need to look more closely at the goals and accomplishments of the ARPA SUR project, the overall current status of speech recognition technology, and the problem areas or "gaps" in technology that remain.

It is useful to consider a "before and after" picture of ARPA SUR impact on continuous speech recognition. When the ARPA SUR project began in 1971, only a few successful laboratory tests had been done on recognition strategies, and almost all of that previous work had been confined to recognition of small vocabularies of isolated words, spoken by one of a few male talkers who had previously trained the system to recognize their voices. Little was known about how to successfully handle continuous speech. The ambitious system specifications of the ARPA SUR project called for machines that would accurately (i.e., for over 90% of

GOAL	RESULTS WITH 1976 ARPA SUR SYSTEMS			
	HARPY	HEARSAY II	HWIM	SDC
Accept continuous speech, ...	184 sentences	22 sentences	124 sentences	54 sentences
from many cooperative speakers, ...	3 male, 2 female	1 male	3 male	1 male
in a quiet room, ...	(computer terminal room)			quiet room
with a good microphone, ...	(inexpensive close-talking mike)			good mike
with slight adjustments for each speaker, ...	20 training sentences	60 training sentences	no training	no training
accepting 1000 words, ...	1011	1011	1097	1000
using an artificial syntax, ...	BF=33	BF=33 or 46	BF=196	BF=105
yielding less than 10% semantic error, ...	5%	9% or 26%	56%	76%
in a few times real time (=300 MIPSS)	28 MIPSS	85 MIPSS	500 MIPSS	92 MIPSS

Figure 3. Goals and final (1976) results for the ARPS SUR systems

the correctly-spoken sentences) accept continuous speech from many cooperative speakers, with near-ideal conditions of quiet rooms and high-fidelity equipment. Sentences were to be highly-stylized structures defined by a small grammar, using a 1000-word vocabulary. Realizing both the complexity of the problem and the prospects for rapid advances in computer technology, they called for the recognition to be accomplished on very large fast computers that could handle about 100 million internal instructions per second (which is about 100 or more times as powerful as the actual computers the systems were finally built on), and yet they allowed the computer processing to take several times as much time as the duration of the spoken sentence. (In computer parlance, the processing then requires "several times real time").

In addition to five original system-building contractors (who tried alternative system designs), the project included four research contractors who were charged with developing advanced ideas for improving the recognition techniques. At the mid term of the project, the best intermediate systems were selected for continued development. The result was that in the fall of 1976, Carnegie-Mellon University demonstrated two alternative system designs (called HARPY and HEARSAY II), Bolt Beranek and Newman demonstrated the "Hear What I Mean" (HWIM) system, and System Development Corporation also demonstrated a system.

The HARPY system developed at Carnegie-Mellon University basically met or exceeded the system goals by correctly understanding 95% of the sentences spoken by five talkers, using a 1011-word vocabulary and a highly-constrained grammar of sentences relevant to a task concerning the retrieval of documents from the computer memory. Five talkers is not "many", and the tests were done on only a

small set of 184 sentences, due to time and money limitations. However, the system did work well even when the original specifications were exceeded by having it handle somewhat noisy speech with inexpensive (lower-fidelity) microphones. Harpy not only met the "letter of the law" by matching the ambitious goals for the project; it also fulfilled the "spirit" of the project by demonstrating the feasibility of a limited (but potentially useful) technology for computer understanding of continuously-spoken sentences. Also, in line with the spirit of the project, it made effective use of strict constraints on allowable (grammatical, meaningful, and relevant) word sequences, to bring the task within manageable limits. Other final ARPA SUR demonstration systems had higher error rates primarily because they dealt with more difficult tasks, used more general techniques that could have been used for additional more ambitious tasks, and were not as carefully tested and adjusted as HARPY before the final demonstration tests.

The successful attainment of the original ambitious system goals was a major contribution of the ARPA SUR project, and the HARPY system performance now provides a baseline or benchmark for assessing future work on continuous speech recognition. However, many other important contributions were made, as is detailed in a recent report (Lea and Shoup, 1979). Most of these valuable contributions are highly technical, but we can make several general observations about ARPA SUR contributions. Of major importance is that we now know that continuous speech can be accurately recognized in the laboratory, at least for the case of sentences related to a limited task. What's more, the original premise that recognition accuracy would be aided by judicious use of linguistic constraints has been vividly demonstrated. The value of artificial intelligence ideas like efficient search strategies and cooperation among several incomplete sources of knowledge has been shown. Several promising alternative system structures have been tested, and major advances were made in certain system components such as vowel and consonant recognition schemes, phonological rules, prosodic analysis routines, and word identification procedures. Thus, the project produced major strides in the necessary technology for commanding machines by naturally spoken sentences.

2.2.4 A Broad Spectrum of Recognition Capabilities - The result from the 26 year history of speech recognition is a variety of different recognition capabilities, as summarized in Figure 4. The easiest task is recognition of isolated words, surrounded by pauses, and then one can use linguistic constraints on allowable sequences of such words, by still maintaining pauses, as in the sequence "...RIGHT ...THIRTY...DEGREES..." When the speech flows freely in connected form, word

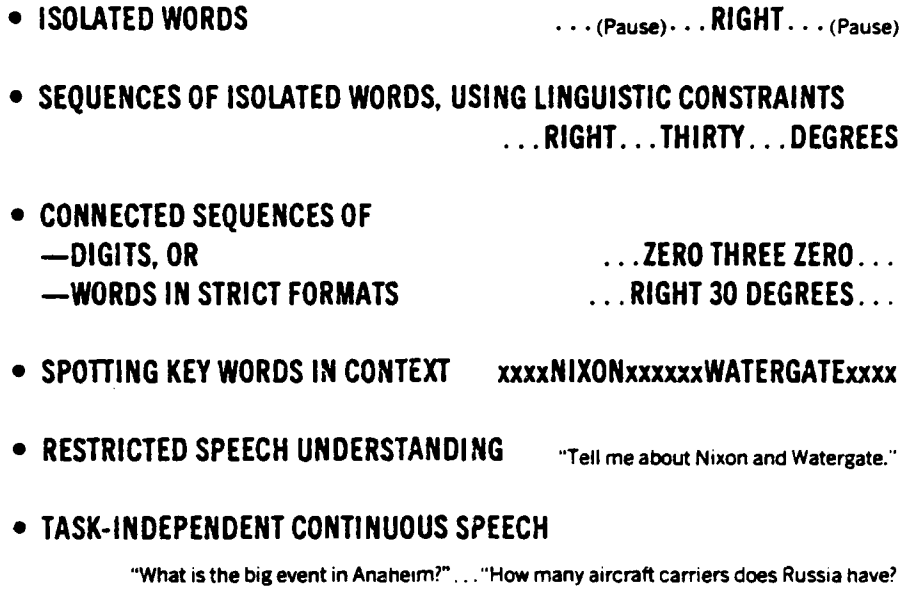


Figure 4. Forms of Speech to be Recognized.

boundaries are hard to find and words distort the pronunciation of the neighboring words, but the task can be limited by handling only strings of digits ("...ZERO THREE ZERO...") or strictly formatted word sequences ("...RIGHT 30 DEGREES..."), where only certain words can go in each slot in the structure. Key information-carrying words can be spotted in the context of conversations, so that one can decide that a conversation is worth listening to based on the occurrence of certain key words like "NIXON" and "WATERGATE." Restricted speech understanding is intended to handle total sentences relevant to a specific task, using all the knowledge that can be brought to bear about the grammatical and plausible sentences one can say when working in a specific task domain. Task-independent continuous speech recognition will allow a change in topic from one utterance to the next and still correctly identify the wording of the sentences, such as illustrated in Figure 4 for the sentences "What is the big event in Anaheim?" and "How many aircraft carriers does Russia have?"

2.3 Current Technology in Speech Recognition

Now, where does all this leave us currently? We have seen that current commercial recognizers handle small vocabularies of isolated words. As shown in Figure 5, current recognizers range over a broad spectrum from cheap hobbyist subsystems to expensive accurate systems ready for fully effective use in field applications. Heuristics, Inc. makes a $299 "Speechlab," which requires a sep-

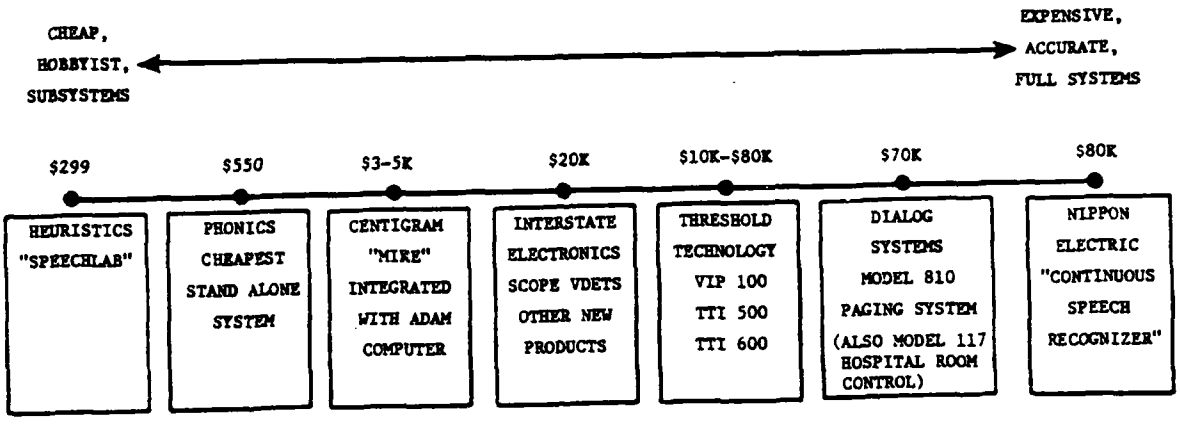


Figure 5. The Spectrum of available commercial speech recognizers.

arate computer, while Phonics has a "stand-alone" system for 16-word recognition, for $550. Centigram's "Mike" system is an order of magnitude more expensive, and even more expensive (but also more accurate) systems are available from Interstate Electronics (who manufactures a version of an earlier recognizer developed by Scope Electronics). Threshold Technology Incorporated offers the leading line of accurate recognizers (including the VIP 100, the TTI 500, and the TTI 600), having sold more than any other systems other than Heuristics' hobbyist device. Dialog Systems Incorporated offers several word recognizers which focus on use over telephone channels, and their Model 810 system sells for about $70,000. Nippon Electronic Company has recently announced a recognizer that can handle either isolated words or connected sequences of digits, using a two-stage dynamic programming algorithm. For two channels, the NEC system sells for about $80,000.

These various recognizers are primarily intended for speaker-adaptive use, requiring a moderate amount of system training for each new talker. Costs of the more advanced systems are expected to come down dramatically in the next few years.

Word recognizers have been effectively used for various hands-busy applications such as:

COMMERCIAL:
- Package sortation systems;
- Inspection and quality control;
- Voice instructions to machine tools (Martin, 1976);
- Voice actuated wheelchairs;
- Hands-free control of hospital room environmental conditions;

MILITARY:
- Cartography (Goodman, et al., 1977);
- Voice authentication systems (Doddington, 1976);

- Training skilled communicators like air traffic controllers (Breaux, 1978; Grady, et al., 1978); and
- Simulations of cockpit communications (Curran, 1978; Coler, et al., 1978).

Current technology also includes industrial development projects for advanced capabilities in speech recognition, as listed in Table I. Notable research work is being done in France, Germany, Italy, and Japan. Also, in Britain and France

TABLE I. CURRENT DEVELOPMENT PROJECTS IN SPEECH RECOGNITION

BELL LABORATORIES	• High-accuracy isolated word recognition with linguistic and task-dictated constraints on sequences, and speaker-independence; • Telephone applications with digit strings and restricted word sequences (e.g., directory assisstance and travel reservations); • Research on entropy and task complexity neasures, speech science, and voice response.
CARNEGIE-MELLON UNIVERSITY	• Minicomputer version of HARPY; • Enhancements of HARPY for bigger tasks, incremental compilation of networks, and automatic knowledge acquisition.
IBM RESEARCH CENTER	• Statistically-based approach to general continuous speech recognition, without restrictive task constraints, but with use of extensive speaker-dependent statistics. Tested with a variety of tasks, including an ambitious 5000-word Lasar Patent Text task. Goal is automatic transcription of unrestricted spoken texts ("dictation machine").
ITT DEFENSE COMMUNICATIONS DIVISION	• Low-cost isolated word recognizers, and speaker independence; • Work spotting in connected speech; • Practical conditions of telephone bandwidth and noisy speech.
LOGICON	• Applications studies with available recognizers, for training air traffic controllers and for other military and training applications; • LISTEN system with Markov model for highly-restricted connected speech recognition.
NIPPON ELECTRIC COMPANY	• Digit string and restricted word sequence recognition; • Advanced dynamic programming methods; • (Other manufacturers are currently working on similar projects).
SPERRY UNIVAC	• Linguistically-based connected speech recognition system; • Word spotting in connected speech, with practical channel conditions.
TEXAS INSTRUMENTS	• All voice talker-verification system with initial recognition of six-digit strings; • Error-correcting methods in digit string recognition.
BRITAIN, FRANCE, GERMANY, ITALY, JAPAN, POLAND	For various studies in isolated word recognition, digit string recognition, connected word sequence recognition, speech understanding systems, and airborne applications, refer to (Haton, 1979), (Lea, 1979a), and (Peckham, 1979).

studies are being conducted on the effective use of speech recognition in operational airborne situations (Bridle and Peckham, 1978; ISPENA, 1978). The commercial interest in speech recognition seems to be expanding rapidly perhaps considerably more rapidly than current military activity in this field.

2.4 Primary "Gaps" in Current Technology

Despite recent important and satisfying advances, the ARPA SUR project and other recent work have demonstrated that in almost every component or aspect of a versatile recognition system, there still is need for further major improvements. The problem of voice input is not solved. No system currently can precisely and correctly identify much more than half of the vowels and consonants in continuous speech, yet experiments conducted during and following the ARPA SUR project suggest that humans can do much better than these current vowel and consonant identification schemes. Recognizers have not even incorporated or adequately tested many of the published rules concerning English phonological structure (the allowable sequences of vowels and consonants, the effects of one sound on its neighbors and vice versa, and the effects at boundaries between words). Prosodic structures (intonation, stress patterns, and timing of speech events) show great promise of aiding word recognition and detection of several aspects of grammatical structure, but prosodic information has had virtually no impact on the performance of previous recognizers. While several promising techniques have previously been developed for identifying words by their resemblances to expected pronunciations, further work is still needed to increase the accuracy of word matching. At the higher levels of linguistic analysis (dealing with larger units like phrases and sentences), efficient constraints must be developed that still allow future expansions to more difficult tasks. We need precise methods for evaluating the total performance of a recognizer.

While current technology offers several commercial devices for isolated word recognition, and one system that is purported to accurately handle restricted continuous speech (digit strings or highly constrained word sequences), still there are major "gaps" in current technology that must be filled before speech recognizers will fulfill their potential as versatile tools for conversing with machines.

In response to a detailed questionnaire about the status of speech recognition technology (Lea and Shoup, 1979, Appendix B), 34 experts with a combined experience of over 300 years work in speech recognition predicted that about 12,000 recognizers could be marketed in each of the next few years, and they ranked the most needed types of systems as shown in Figure 6. (A marketing consultant predicts

1. Rank order the types of speech recognition or understanding that are most needed now:

A	N	T	
2.55	3.88	3.36	Very low cost ($_____) isolated word recognizers
0.91	3.35	2.39	Moderate cost ($_____) isolated word recognizers
0.55	2.24	1.57	Larger-vocabulary (_____-word) isolated word recognizers
0.91	1.59	1.32	Recognizers that handle sequences of words, where successive words are separated by short pauses, without use of higher-level constraints
2.27	4.71	3.75	Recognizers that handle sequences of words separated by short pauses, *with use of higher-level* (syntactic, semantic) constraints
2.18	5.36	4.11	Digit string recognizers
4.0	3.88	3.93	Recognizers of connected word sequences that are highly constrained to certain formats
2.91	3.88	3.50	Recognizers of modestly-constrained word sequences
1.82	1.29	1.47	Speech understanding systems that are much more restricted than HARPY
3.64	1.76	2.50	Fairly-well-restricted (HARPY-like) sentence understanding systems
1.55	2.0	1.82	Speech understanding systems that are somewhat less restricted than HARPY
3.18	0.53	1.57	Speech understanding *systems that are much more* powerful (unrestricted) than HARPY

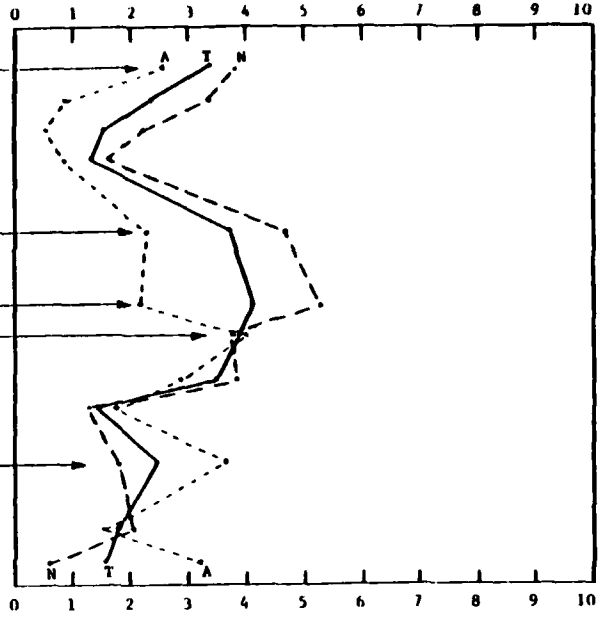


Figure 6. This page from the questionnaire by Lea and Shoup (1979) shows the summary of responses by 34 experts to the question of which types of speech recognition or understanding systems are most needed now. The numbers on the left represent average ranks of relative need, with a maximum possible of 10, for total agreement about the most needed system.

The column T is a total rank for all respondents, and the A and N columns provide rank scores for the respondents who were ARPA SUR participants and nonparticipants, respectively. The corresponding plots of relative ranks are shown on the right, with the total (T) scores showing digit string recognizers as most needed. Arrows point out several most needed types of systems.

an expanding ten-year market of over $1.5 billion; Nye, 1979). A whole spectrum of recognition capabilities are included in this list of most-needed systems. The experts have thus called for systems of various complexities to fill the long-standing gap between practical isolated word recognizers and sophisticated but impractical laboratory work on continuous speech. In the recent survey, Lea and Shoup also found widespread agreement about other primary "gaps" in current technology. Major work is needed in improving the acoustic-phonetic "front end" of a recognizer, along with providing new capabilities in prosodic analysis and phonological rules. Performance evaluation is another area that needs further work, so that total systems can be adequately tested, and sources of error or inadequacy can be determined, with the performance of each component in a system also clearly understood. Methods are needed for measuring the complexity of tasks, so we can decide if one system that gets 90% recognition accuracy on a simple task is truly better or worse than another system which yields 50% recognition on a difficult task. Methods for assuring near real-time processing and uniform scoring of hypotheses are also needed. Extensive basic research is also needed on such topics as acoustic parameters and distance measures, phonetic word boundary effects, coarticulation, and acoustic phonetic and prosodic characteristics of English sentences. Detailed lists of gaps in current technology, and their relative priorities, are listed by Lea and Shoup (1979, Figure 4-1), and are worthy of the reader's careful scrutiny.

2.5 General Recommendations for Advancing Speech Recognition

Lea and Shoup (1979) offered general recommendations for advancing speech recognition, by "bridging the gaps" in current technology with the various project types and problem areas listed in Table 2. At least four types of speech recognition projects are needed now, including: (1) applications studies with available commercial recognizers; (2) evaluations of available devices and specific improvements (without major re-design of systems); (3) advanced development projects to substantially expand recognition capabilities; and (4) research on necessary knowledge sources and basic concepts relevant to future success in recognition.

To fill the gaps in the longer-range research and advanced development aspects of current technology, one might conceive of another large project like the ARPA SUR project. Experts polled by Lea and Shoup (1979, Appendix B) were of the opinion that if such an ARPA SUR-like project were attempted again, it should not have a fixed set of system specifications and deadlines, but should be an open ended project directed at several systems representing a spectrum of task

TABLE 2. RECOMMENDED TYPES OF SPEECH RECOGNITION PROJECTS

APPLYING AVAILABLE RECOGNIZERS	• Cockpit Communications (NASA Ames, NADC, NTEC, RADC) • Training Air Traffic Controllers (NTEC) • All-Voice Access Control to Secure Areas (RADC) • Voice Entry of Cartographic Data (RADC, DMA) • Key-Word Spotting (RADC, ONR) • Computer-Assisted Trouble Shooting
EVALUATING AND ADVANCING CURRENT TECHNOLOGY WITHOUT MAJOR RE-DESIGN OF SYSTEMS	• Comparative Evaluations of Alternative Input Modalities • Comparative Evaluations of Alternative Speech Recognizers Selection of databases and benchmark tasks Measuring complexities of tasks • Human Factors Studies Effects of physical and mental stress Design criteria for user acceptance • Realistic Channel Conditions Microphone characteristics and placement; telephone input; Noise; channel distortions • Larger Vocabularies • Application of Isolated Word Recognizers to Word Spotting • Speaker Independence without Extensive Training
DEVELOPING ADVANCED SYSTEMS	• Evaluating the Need for Continuous Speech • Digit String Recognizers • Word Sequence Recognizers • Moderately-Restricted Speech Understanding Systems • Autonomous Continuous Speech Recognizers • Methods for Fast Processing of Extensive Data
RESEARCH ON NECESSARY CONCEPTS AND KNOWLEDGE SOURCES	• Acoustic Phonetic Analysis • Prosodic Aids to Recognition • Performance Evaluation and Task Complexities • Phonological Rules • Linguistic Constraints on Ambiguities • Scoring Procedures for Selecting Hypotheses • Social Issues Concerning Uses of Recognizers

complexities, and should use practical input transducers like the telephone, while demanding accuracy of over 95 to 99% on either sentences with a several hundred word vocabulary, or constrained and formatted word sequences.

While another large-scale project of cooperating contractors under one sponsorship would be possible and reasonable, other alternative mechanisms exist for bridging the research and development gaps in current technology. Cooperative effort among several military service groups might be possible through the recently-established Technical Advisory Group ("TAG") on voice interactions. The establishment of two or three "speech science centers" would be recommended, to bring together speech and linguistic expertise, powerful computer facilities and computer network capabilities, visiting researchers who use such facilities to advance their work on aspects of recognition, etc. Such centers could compile valuable databases of spoken utterances for analysis, plus lists of publications on various related topics, and a variety of computer programs for various aspects of recognition. See (Lea and Shoup, 1979) for more detailed discussions of needed research and development efforts.

On a more practical vein, the immediately foreseeable applications of recognition would suggest further focus on improved recognizers of isolated digit strings, and formatted word sequences. The top fourth of [illegible] military applications that seem appropriate for available word recognizers [illegible] agencies shown in parentheses have already funded or expressed interest in the applications as listed. Rome Air Development Center and the Defense Mapping Agency have already demonstrated the value of voice input of heights and depths at coordinates on a terrain map or oceanographic map, but further work in actual field use is called for, and it looks like connected digit recognition might help this application. Logicon and the Naval Training Equipment Center have developed an excellent application in the training of air traffic controllers for ground controlled approach. The skills being learned are primarily verbal, and controllers must learn not to deviate from agreed-upon short utterances. Automation of instructor and pilot simulation functions is possible when the trainee's utterances are automatically recognized and corrected by machine responses. There may be many other such verbal training applications, which could prove as appropriate for speech recognizers as package sorting, inspection, and machine control have proven to be in commercial applications.

The cockpit communications applications will be discussed in sections 3 and 4. The RADC effort in having Texas Instruments develop an all-voice access control system for base internal security purposes is another application that deserves further work.

One promising but apparently untapped application for available recognizers is in computer assisted trouble shooting. An apprentice trouble shooter usually asks specific questions and communicates simple data to his more-knowledgeable supervisor, and such guided testing could be accomplished by interactions with a computer-stored trouble shooting information source.

The expert respondents to the Lea and Shoup survey agreed with the importance and appropriateness of each of these previously-mentioned applications. However, they did not agree with either speech input of commands to gunfire control computers or key word spotting. Key word spotting has been the subject of considerable previous research and development projects, and still seems to be one of the simplest extensions of recognition capabilities to continuous speech. While its application to automatic surveillance of communication systems might be controversial, it can also serve as a word hypothesization process of a speech understanding system. Also, word spotting might be used in the captioning of television programs as an aid to the deaf.

In the next section of Table 2, I list some of the projects needed to comparatively evaluate recognizers and extend their current capabilities. While a few studies have been done to compare speech input with other modalities of communication with a computer (e.g., Ochsman and Chapanis, 1974; Welch, 1977), more human factors studies are still needed, particularly with realistic situations using actual computer input devices.

With the advent of many commercial sources of word recognizers has come a growing demand for objective procedures for evaluating systems. Several manufacturers have endorsed the need for industry standards, such as general speech databases of 100 or more talkers speaking many instances of the digits and other vocabularies. Also, recognizers need to be comparatively evaluated with realistic "benchmark" tasks of interaction between unsophisticated users and voice entry systems. In addition, to facilitate evaluations without always using the same task, it would be advantageous to develop good measures of task complexities, so one can compare 90% accuracy on any easy task with 60% accuracy on a difficult one. Other aspects of recognizer evaluation relate to various human factors studies and the effects of physical and mental stress on performance of the recognizer and human user. Systematic studies should be done to determine what accuracy and other system features are needed for user acceptance of the recognizer.

Tests are needed with various microphone characteristics and placements, telephone input, noisy environment, and various channel distortions.

Finally, some straight-forward extensions of recognizer capabilities should be considered that do not require major system re-design or advanced development

projects, so that larger vocabularies, spotting of words in context, and speaker independence without extensive re-training will be possible.

Some aspects of the advanced development projects shown in the third portion of Table 2 deserve consideration by those interested in practical recognizers to be used in operational environments. A large gap exists between currently available (and usable) accurate isolated word recognizers, and long range work on ambitious "speech understanding systems". Most practical applications of speech recognition have been demanding "off-the-shelf" systems, and thus have been trying to adapt their requirements to isolated word recognizers. No absolute needs for continuous speech recognizers have been defined, at least not in the military (Beek, et. al., 1977), but this may be due in part to the available technology dictating to the application rather than allowing the applications to determine needed technology. Studies should be done to determine when it is truly profitable to use connected speech, and if so, how much language versatility is really useful for each application. What is the tradeoff between added complexity, costs, and potentially higher error rates introduced by use of connected speech, versus simplicity, presumably lower costs, reliability, and a long history of field testing that are associated with isolated word recognizers?

Carefully limited connected speech recognizers are now on the threshold of practicality. Digit string recognizers and word sequence recognizers need to be refined substantially so that they provide accuracies comparable to those of word recognizers. Speaker independence without the need for extensive training must also be assured. Indeed, for many commercial applications (e.g., in telephone banking and credit transactions), as well as military telecommunications operations, speaker independence is probably more important than making inroads into the problems of connected speech recognition. For all types of continuous speech recognizers, work seems appropriate to assure fast processing so that systems respond in real time and also so systems can be quickly developed and tested with extensive speech data. Such speed considerations could be integral parts of other recognizer developments, but the importance of fast processing warrants explicit mention, and some work might even be appropriate on system-independent developments of fast processors. A related topic of practical concern is the development of low-cost micro-miniaturized recognition hardware (large scale integrated circuits, "computers-on-a-chip", etc.), which will bring speech recognition into the practical commercial domain for large markets, and which will allow extremely small speech input devices for cockpits and related cramped spaces. Such low-cost micro-miniturization will require a large initial investment such as can only be provided by very large semiconductor or computer corporations, or by substantial governmental

funding that permits mass production of speech processing "chips".

In summary, the field of speech recognition deserves stable funding, excellent facilities, and speech science centers where advances can readily be made and monitored. The field also needs systematic studies of applications, and judicious improvements in limited speech recognizers. Careful attention must be given to transferring the on-going results from the recognition development projects and research into various practical DOD applications. A coordinated program of further DOD work is needed, is possible, and deserves immediate attention.

3. STUDIES OF AIRBORNE USES FOR SPEECH RECOGNIZERS

The expanding history of successes in speech recognition, the current state of technology, and the potential benefits from the future work called for in section 3, all suggest a bright future for voice input to machines. Yet, before speech recognition facilities are incorporated into airborne systems, some remaining issues need to be considered. General questions arise, such as: Why should speech recognition be attempted in cockpit situations? What are the advantages and disadvantages of airborne applications of speech recognizers? What peculiar problems does the cockpit application introduce? What tasks of airborne crews are particularly suitable for the application of speech recognition technology? What studies have been done, or are currently being pursued, to firmly establish the capabilities and dimensions of difficulty in voice input to airborne machines? What problems and issues remain, and what are the relative significances of these remaining issues?

In this section, we will attempt to outline the advantages and disadvantages of voice input in cockpit situations (section 3.1), the previous studies that have explored key topics in airborne applications (section 3.2), and some of the current (1979) work being conducted throughout the USA and other NATO countries (section 3.3). Remaining issues that deserve further consideration will be considered in section 4. Recommendations for further work will be outlined in section 5.

3.1 Advantages and Disadvantages of Cockpit Applications

Knowledgeable workers in the field of speech recognition will often quickly point out several disadvantages to the use of voice input for machines aboard aircraft. Each of the following dimensions of difficulty pose questions about the utility of airborne speech recognizers:

Noise: sometimes environmental noise is louder than the voice;

Vibration: voice quality degrades and input conditions rapidly vary;

G forces (Acceleration): the microphone moves, the mouth distorts, and the speaker is strained;

Oxygen mask: the voice is somewhat unnatural, and microphone placement is not fixed;

Speaker variability: crewmembers cannot be selected on the basis of dialect or consistent enunciation:

Stress and fatigue: overloaded crew members tire, and speak under emotion or duress;

Size and weight constraints: devices must be as small and light as possible;

Training requirements: new training techniques, crew performance evaluations, and speaker-dependent adjustments of the systems must be incorporated;

User acceptance: highly skilled pilots and crew members are not quick to accept novel technology or untested devices;

Criticality of errors: navigation, communication, and armament actions must not be mistakenly initiated;

Limited market: research, development, and production costs must be borne by a limited number of operational systems; and

Costs: speech recognition systems currently cost more than keyboards or other tactile input devices.

However, it should also be understood that voice input has a number of distinct advantages in its favor within the constraints of airborne applications. To begin with, there are the basic benefits that the voice modality generally offers, but which are enhanced in airborne situations:

Naturalness: spoken commands may exhibit fewer substitutionary errors, and errors in transposition of successive commands; also natural communication is probably more reliable and spontaneous than unnatural modalities, under stress or extenuating circumstances like weightlessness, g forces, vibration, fatigue, etc.;

Fast, high-capacity communication: speech is faster than the alternative input modes for complex tasks, and offers potentially prompt commanding of machines in critical phases of missions;

Frees hands and eyes for other tasks: the visual and tactile channels of crew members are even more overloaded than in many commercial applications, where speech recognizers have already proven very effective;

Permits multimodal communication: some tasks can be done with currently used keyboards or controls while other actions are simultaneously performed by voice command;

Possible at various orientations and distances, in darkness, and around obstacles: crew members can speak commands while turning to read displays, studying terrain and extra-vehicular situations, adjusting distant controls, or working in the dark conditions of night flight or blinding light after armament blasts.

Flight of complex airplanes and helicopters introduces crucial cases of overloaded workers, with busy hands and eyes. The visual and tactile channels are already approaching overload from too many displays, keyboards, controls, and related devices, all of which compete for the crew member's attention and require valuable cockpit space. We can partly resolve the problem of excessive workload by transferring visual and/or motor tasks to the less saturated vocal and auditory channels of the crew member. Such unburdening of the operator is particularly valuable during those portions of missions during which workload is especially high. Speaking or listening does not need to disrupt attention or interrupt actions on other critical tasks, such as is required when one turns for a line-of-sight view of displays or moves into position before a keyboard or control device.

As we shall see in more detail in section 3.2, speech has been found to be faster than other modalities for commanding machines during some (but not all) tasks. The tasks for which speech is particularly effective are the more complex tasks, especially when the human is occupied at other tasks involving hands and eyes. There are many such tasks in airborne situations.

In addition to introducing these cases of accenting the general advantages of voice input, the cockpit situation introduces several other distinct features that constrain the task of speech recognition:

Limited speaker population: the recognizer need handle only the few members of the airborne crew, not arbitrary speakers;

Highly trained speakers: the crew members are extensively trained and cooperative speakers;

Microphones are already in use: "mike fright" and related human factors are minimized, no new input device need be dealt with by the human, and the speakers are trained in proper microphone usage; and

Restricted tasks and command sequences: small vocabularies of alternative commands are allowed, with strict syntax of command sequences usually already dictated, and tasks already defined in minute detail.

Airborne situations do not have the complication introduced by allowing any arbitrary speaker to walk up to a microphone or pick up a telephone to talk with a speech recognizer. Crew members are already trained to produce standard, consistent communications. It is possible to train the system to a crew member's voice at the same time that he or she is being trained to effectively use the airborne systems. The great potential for using speech recognition in training applications has already been well acknowledged (Breaux, 1978; Coler, et al. 1978; Curran, 1978; Grady, et al., 1978; Feuge and Geer, 1978), and hinges in good part upon the fact that current skills being learned are primarily verbal communication skills (as with airtraffic control, airborne communications, etc.). Training uses of speech recognizers can be readily translated into airborne systems to permit keeping a running account of user and airborne team performances.

Coler and his colleagues at the NASA Ames Research Center have specifically discussed the potential of flight applications of speech recognition systems, and have noted specific aspects of airborne situations that make speech recognition appear particularly attractive. They noted (Coler, et al., 1978, p. 164) that speech input to airborne systems could:

"(1) reduce the difficulty (and risk) of performing high-workload manual control procedures during critical phases of flight,

(2) minimize eye-hand coordination problems that often accompany a heavy manual control burden; and

(3) reduce visual time-sharing requirements so that attention can remain focused outside the cockpit or upon a primary flight display for longer periods of time."

They specifically noted the selection and tuning of radio frequencies as a procedure that often interferes with other essential manual tasks, and observed that computer generated displays could be voice controlled, to allow visual attention to remain focused on the display. They suggested that whenever command sequences must be performed manually in integration with other manual tasks and must be executed rapidly, voice commands have the potential for

providing faster and more accurate performance that is less disruptive of primary tasks.

Both the NASA Ames studies and the NADC/ONR sponsored work (including the study by Boeing/Logicon; Feuge and Geer, 1978) have focused on the Navy P-3C Orion anti-submarine aircraft. (NASA Ames has also considered helicopter applications.) NASA Ames workers specifically noted the low altitude flight situations when the P-3C pilot must select and execute command functions using a 35-key keyset located slightly behind him to the right. The pilot must turn his body towards the keyset, directing visual attention away from the outside visual scene and primary flight instruments. Voice input offers a promising alternative. They also noted that, during some missions, the desired rate of information entry in the Sensor Station 1 and 2 crew positions of P-3C greatly exceeds the rate obtainable with current keysets and thumb wheel switches. As the role of human-to-machine interactions continues to expand with onboard avionics systems, crew members must be able to provide inputs in a manner that is "(1) accurate, (2) tolerant of errors and updates, (3) rapid enough to meet the demands of the task at hand, (4) natural and convenient, so that use of the input system does not add significantly to the user's workload, and (5) interruptable." We must explore whether speech input has the potential of exceeding conventional entry modes in these dimensions, within the airborne environment.

Part of the answer as to how speech input technology compares with conventional input/output devices is to be found in the studies by Welch (1977; also Martin and Welch, 1979). Using <u>available commercial devices</u> (not theoretical speculations or human-human interactive simulations of human-machine interactions), he investigated the speed and accuracy of isolated spoken word interactions, keyboard entries, and graphical (menu) data entry systems. With the simple data entry task of copying numeric strings, the keyboard was considerably faster and more accurate than voice. The keyboard was also somewhat faster than voice in a simple alphanumeric string entry task. However, voice entry provided the <u>lowest error rate</u> (i.e., most accuracy) for the simple alphanumeric scenario. This was apparently due primarily to the greater immunity to reading errors when voice was used. Quite significantly, when the scenario of data input was more complex, involving entry of complex <u>flight data</u>, voice entry was <u>faster</u> than keyboard entry, for inexperienced subjects, and had a similar operational error rate. However, it should be pointed out that the commercial speech recognizer's error rate <u>before corrections</u> was higher than that for keyboard entry with the complex flight data; it was only the operational method for error correction that saved voice from having a higher error rate. Still, it is significant that even

with corrections being made, the voice mode was faster for the complex task. Also of primary significance in airborne applications is the result that voice input had an important speed advantage when the hands were occupied in other tasks. Graphical entry was never the most accurate or fastest input mode.

One might conclude that simple entry of numerical data might best be accomplished with keyboard, particularly for situations where experienced operators are involved, and the hands are not occupied in other tasks. Voice input seems more appropriate whenever the hands are occupied, the entry task is complex, or the operators are inexperienced. It is significant that the promising performance of voice input in this study was based on a highly-restricted form of voice entry; namely, isolated words, recognized by a commercial device of necessarily limited accuracy. Even better results might be predicted for connected speech and improved recognizers. There seems to be a clear potential for improved airborne crew station performance wherever the entry task is complex or the hands and attention are occupied in other tasks. As we will see in sections 3.2 and 3.3, recent and planned studies are directed at specifically demonstrating this promising role of voice input in airborne situations.

One of the primary remaining questions concerns how well the speech recognizer will perform in airborne situations. Another concern is determining what airborne tasks are particularly suitable for incorporation of this new technology. Part of the answers to such questions are indicated by previous studies, which we shall now consider.

3.2 Previous Studies of the Cockpit Situation

No experience is available regarding actual airborne applications of speech recognizers, and only a few studies have been done with simulated flight conditions. Other work has been in the form of "at the desk" conjectures and laboratory experiments, plus explorations of crew station tasks that might be amenable to introduction of voice entry devices.

In 1977, Montague of SCOPE Electronics Incorporated reported to RADC on a study with the SCOPE Voice Data Entry System (VDETS), involving isolated word recognition with speech obtained from a centrifuge. After obtaining initially poor performance with the face mask and various levels of g-force, the SCOPE workers modified the VDETS algorithms to increase the coding resolution of the data, to improve the segmentation techniques, to permit multimode training procedures, and to eliminate breath noise. Even without g-forces applied, the recognizer's accuracy (with the face mask microphone) dropped from the usual laboratory performance level. While some variations (reduced accuracy) with

g-forces were found, no consistent pattern could be discerned, and it appeared that the face mask was one of the primary sources of the problems. It appears that the simplest, most accurate summary of this study would be that the SCOPE VDETS did not work adequately with face mask and g-forces, and that some straightforward modifications or enhancements of VDETS algorithms were not sufficient to attain adequate performance under such simulated airborne conditions.

Other studies also suggest that the high accuracy attained under laboratory conditions is often not experienced in the less-ideal conditions of an operational environment (Herscher, 1978; Martin, 1976; Scott, 1977; Breaux, 1978; Curran, 1978). For example, Breaux (1978) found that speakers were able to experience 94% correct recognition with a Threshold Technology VIP-100 recognizer after several hours of "introduction" to the device in the laboratory. Yet, when these students in the Navy's Air Traffic Control School were exposed to "free runs" in which they had full (simulated) control over the aircraft, recognition accuracy suffered. Hesitations, repetition, and corrections were made. The primary difference in these two situations was between speaking commands as fully dictated by the system promptings (that is, "read speech") versus human initiation of all commands ("spontaneous speech"). Similarly, studies have shown significant (sometimes drastic) reductions in accuracy due to operational conditions like noise, restriction to telephone bandwidth, and spontaneous operational speech versus carefully articulated training utterances.

Dr. Robert J. Wherry, Jr., of the Naval Air Development Center, studied the effects of vibration, g-forces, oxygen mask, mission duration, and cockpit temperature on voice quality (as assessed by speech recognition accuracy). This study of the effects of simulated flight was done on the NADC human centrifuge, using a SCOPE Electronics voice command system (VCS) and a Votrax VS-5 speech synthesizer. He found that: (1) voice quality degrades after 0.5 hours with an oxygen mask; (2) voice quality degrades under high (0.3g) vibration; and (3) voice quality degrades under high levels of g-force. The degradation under high acceleration may be attributable to slippage of the mask. Wherry developed a syntactic handler that permitted medium size vocabularies (250 words) and highly flexible statement formats (with alternative ways of saying specific messages). Later, NADC developed and documented a transportable FORTRAN version of this Voice Recognition and Synthesis (VRAS) system.

Curran (1978, p. 130) noted that, before an advanced development program can be initiated and pursued to test and use speech recognition technology in an airborne application, high-payoff airborne uses of voice interactive systems

need to be identified, a methodology must be developed for assessing technical feasibility of voice technology for each proposed application, and mutually supportive basic research and advanced development projects need to be defined into a cohesive program plan.

Government applications and specific crew station tasks amenable to voice input have been explored for NADC and ONR by Boeing and Logicon, and a program plan has been defined (Feuge and Geer, 1978). This is the only known previous comprehensive attempt to identity applications for speech recognition in specific airborne situations, and deserves careful scrutiny. The Boeing/Logicon effort involved reviewing available manuals for the operation of the P-3C anti-submarine aircraft, to identify, for each crew station, those tasks that were suitable for voice technology. Four factors were rated for evaluating each task, then the four factors were combined into a single digit indicating payoff from speech technology, and finally this digit was modified based on the criticality of the task (e.g., could it cause fatal errors in system operation or mission effectiveness) and the frequency with which such a task occurs. Since speech recognizers still make errors, early use of speech technology should not involve critical aspects of flight, deploying armament, etc.

The four factors involved in the Boeing/Logicon rating included:

(a) the technical feasibility of implementing voice to accomplish the task;
(b) the utility of implementing voice;
(c) time and accuracy requirements for the task; and
(d) the impact of unassessed variables such as noise and mission duration.

The technical feasibility was weighted very high in determining the Boeing/Logicon rating, so that it was impossible for a task to be rated as having high pay off for voice input unless it was amenable to use of current (1978) limited-vocabulary, speaker-dependent isolated word recognition technology. In addition, top priority tasks for voice had to have the highest projected utility, in that voice clearly would benefit the crew member for that task (not just match current system capabilities). They assigned a very cautious level of required recognition accuracy, asserting that unless the task could allow a 20% recognition error rate (i.e., only 80% recognition accuracy), the task could not be rated as one for which high payoff was to be possible from voice input. Any task which required over 90% recognition accuracy was categorized as of "questionable pay-off" (at best). "Some pay off" could involve a required moderate accuracy of 90% only if the

utility of voice in that task was clearly high. Critical tasks of frequent occurrence during a mission were then diminished further in payoff rating, unless they were amenable to current technology, of high utility, and did not require accuracies above 80%. This made it particularly difficult for a task to get a high rating for introduction of voice.

One result of this analysis was that the less-critical Sensor 1 and Sensor 2 crew stations were concluded to be the most amenable to introduction of voice input and output technology. The pilot position was next most appropriate. Thus, some non-critical tasks of monitoring indicators, activating switches, entering data, and adjusting some controls were included among those most amenable to introduction of speech recognition and synthesis. Airborne subsystems that showed promise of productive use of speech technology included (in decreasing order, for the pilot station) communications, propulsion, search stores, photo, and data handling.

Based on these studies, Boeing/Logicon proposed 28 projects (later reduced to 25 projects) which might form a part of a comprehensive program plan for introducing voice technology into P-3C (or similar) crew station operations, or into related applications of training air crews and measuring crew performance. They numerically rated and rank ordered the projects, considering: (1) the estimated impact, or advantage of the project results to future users; (2) the risk of project success, including required time delay, accuracy rates, background noise, and other variables; and (3) the cost, including facilities, equipment, personnel , and time required for completion and evaluation. Three separate subjective judgments or ratings were done by individuals working in the Boeing/Logicon study.

As Curran (1978) observed, this Boeing/Logicon study was not intended to absolutely identify the best voice applications for the P-3C or to strictly define needed projects, but merely to develop methodologies required to identify high payoff applications of voice technology. While the Boeing/Logicon study may help guide the NADC advanced development program in Voice Interactive Systems Technology (VIST), the Program Plan which they recommend differs somewhat from that to be proposed in section 5 of this report.

In section 4, we shall consider some recommendations for further work that may broaden and strengthen the overall program for applying speech recognizers to airborne tasks. We shall also consider some previous work and recent technology trends that might alter certain aspects of the Boeing/Logicon program plan. These suggestions and the resulting recommendations do not, however, diminish the value of the general Boeing/Logicon methodology used in: analyzing

crew station tasks; rating them on the basis of technical feasibility, utility, time and accuracy requirements, criticality, and frequency of performing the tasks; and identifying top priority tasks for incorporation of voice technology. From such applications analyses and a comprehensive understanding of current and forthcoming technology, one can define a program plan of needed projects.

Another study of airborne applications of speech recognizers, at the NASA Ames Research Center, is also considering the P-3C crew stations (and helicopter pilot tasks) as representative systems. The NASA Ames project began in 1972 with the use of a SCOPE Electronics recognizer. After limited accuracy was attained with the initial device, new software programs were implemented to improve the accuracy to over 99% for the digits (Coler, et al., 1978). Perhaps one of the most significant aspects of the NASA Ames work concerns their study of the effects of vocabulary size and word confusions on the accuracy of recognition. They constructed a 100-command flight vocabulary for full mission simulation. Each of 10 untrained male speakers provided 100 test utterances of each of the 100 commands, for a total of 100,000 test utterances, which is a very large database. The command language syntax allowed only a few of the commands at any given time during a mission, so the 100 commands were grouped into 15 subsets ranging in size from 3 to 10 commands. Recognition accuracy for the entire vocabulary when syntactic constraints were ignored (that is, when any of the 100 words was considered as possibly spoken) was 95.7% correct, with 5% of the utterances rejected (that is, 5% considered too difficult to identify due to confusions with similar words). When subsetting was allowed, 99.6% of the digits were correctly recognized, with the same 5% rejections. Thus, errors reduced as the vocabulary size was reduced.

Figure 7 graphically shows the error rate in recognition plotted versus vocabulary size, for each NASA Ames sub-vocabulary and the total vocabulary. Though the results are scattered somewhat, there is a general trend towards increasing error rate with increase in the size of the vocabulary. The NASA workers also found that the digits (zero through nine) comprise a relatively difficult 10-word vocabulary, and appear to be an excellent small vocabulary for evaluating a speech recognition for flight system use. They found the digit "six" to be one of the best-recognized words, while the digit five was the most difficult word to recognize, so the digit subset covered nearly the entire range of word-identification results over the 100-command vocabulary. They suggested that:

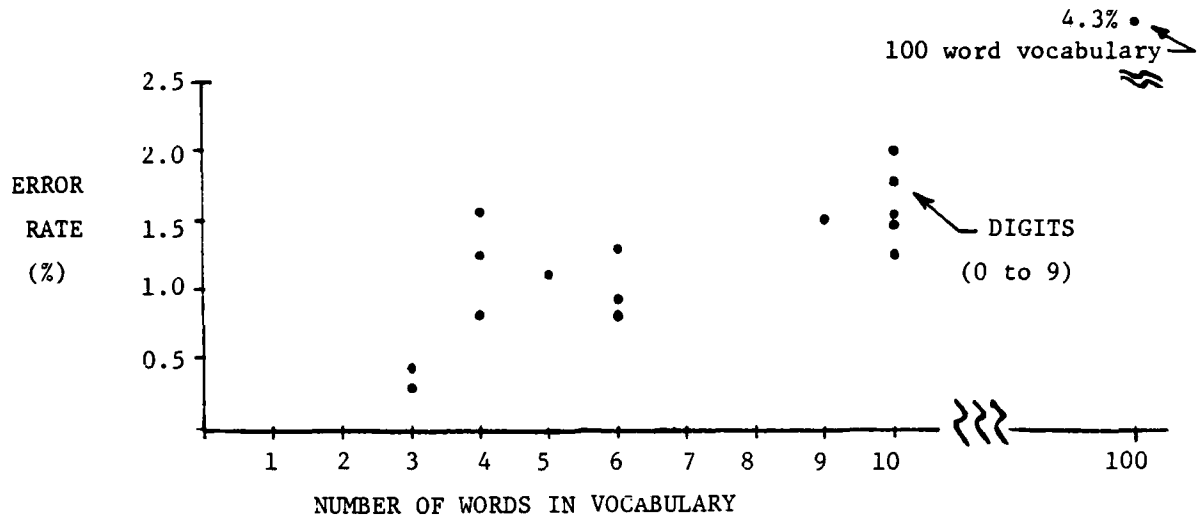


Figure 7. Effects of Vocabulary Size on Recognition Error Rate.

> "New vocabularies proposed for flight systems use may be evaluated by the process described for evaluation of the 100-command vocabulary. Thus any serious incompatibilities between commands within a subset may be identified and corrected, and recognition accuracy for the entire vocabulary can be determined prior to actual use in a flight system". (Coler, et al., 1978, p. 159).

One important aspect of any development of an airborne application of speech is the selection of a suitable vocabulary, as evidenced by the NASA Ames study, and by the inclusion of a "Vocabulary Development" project as seventh in priority in the Boeing/Logicon program plan. However, vocabulary size, and applicability of the vocabulary to any specific airborne task, are not the only factors to be considered in selecting a suitable vocabulary. The confusability of words in the vocabulary, or within the subvocabularies, must also be considered. Some equal-size vocabularies in the NASA Ames results summarized in Figure 7 have different error rates, primarily because of varying amounts of confusion (or similarity of pronunciation) between words within the subvocabularies. We shall consider in section 5 some general studies that can be done to aid in vocabulary selection, giving consideration to the confusability and size of each subvocabulary, the possible syntaxes of command sequences, and the desire to have a straightforward and systematic procedure for selecting sub-vocabularies for a variety of airborne tasks.

3.3 Current Interests and Projects for Airborne Applications

The NASA Ames study is currently being extended to include evaluation of recognition accuracy for test subjects performing a single-axis compensatory

tracking task while exposed to several different conditions of noise and vibration. Four noise conditions will be tested: no noise, helicopter noise at either 90 or 100db, and random ("white") noise at 100 db. The four vibrations to be simulated and tested include: no vibration, smooth and rough jet transport cruises, and helicopter cruise. Tracking errors, and the accuracy and speed of voice and keyboard data entries will be monitored. The NASA work also includes consideration of ruggedized hardware.

Other work is currently being undertaken in several NATO countries. At the Royal Aircraft Establishment (RAE) in Great Britain (UK), a program is underway to evaluate voice entry in cockpit operations, and to determine which tasks or functions will be most effectively helped by voice input. To determine how a pilot might best use a hypothetical perfect recognition device, a human has been used to interpret the voice commands spoken by subjects who were performing a compensatory tracking task. Subjects were prompted (over earphones) to enter 6-digit commands by keyboard or voice. Under a baseline condition without the tracking task, keyboard and voice entry were about equally rapid and accurate. When occupied with the compensatory tracking task, voice entry accuracy improved while keyboard entry accuracy degraded, with little difference in speeds of the two modalities. The accuracy of tracking itself was less degraded by the additional task of voice entry than by the additional task of keyboard entry. RAE is also studying the effects of stress and aircraft noise on voice entry, the effects of mask microphones and improperly positioned microphones, the performance of a Threshold Technology 500-S system in a cockpit simulator, etc.

In France, the ISPENA consortium is planning to use a version of the LIMSI recognition system (cf. Haton, 1979) to help evaluate voice in an avionics control environment, using tapes recorded under various flight conditions in a Caravelle aircraft. As with USA and UK studies, the ISPENA study involves comparisons of voice, a head-up display with menu-selection by trackball and pushbutton, and other data entry methods. In agreement with the Boeing/Logicon/NADC and UK studies, the ISPENA study suggests that critical functions should not yet be considered for voice control, but that radio setup, navigation data entry, and weapons preparation should be appropriate for voice input. ISPENA studies currently include definition of an appropriate 130-word vocabulary and command syntax yielding typical branching factors of 6 to 10, for sequences of isolated words. The recently improved LIMSI system is reported to yield about 100% accuracy on numbers, up to 92% on spoken letters (A,B,C, etc.) and about 100% on the "Alpha Bravo..." phonetic alphabet.

In the USA, continued interest exists at Rome Air Development Center in

developing adequate procedures and evaluations of voice input for cockpit applications, though (as with the French and UK studies) no specific aircraft has been selected for study. All of the cooperating NATO projects include among their top priority avionics studies specific concern for: operation in very bad environmental conditions (noise, distortions, vibrations, acceleration, etc.); small size and weight; limited forms of connected speech (digit strings and other formatted connected word sequences) as well as isolated words with task syntax and subvocabulary selections; and careful study of the effectiveness of voice for each specific task. They also are concerned with the ability to set up for new speakers without requiring that each word in the total vocabulary be uttered as training data (i.e., reduced speaker dependency and reduced training demands) and also the ability of the speaker to adapt himself to the system based on previous successes and errors in dialogue.

The NADC program for airborne applications of voice technology (Curran, 1978) seems to be at the forefront of current studies. Based on the prior exploration of tasks that are suitable for voice input, plus the development of prototype tests under airborne conditions, NADC plans to design, develop, and test a voice system and simulator for selected airborne tasks, with provisions for monitoring and evaluating interactions between operators and the voice system. Proposed task applications will be evaluated for cost effectiveness, contribution to total system effectiveness, and operational acceptability. The result will be detailed system specifications for implementation of voice applications.

The five year development effort at NADC includes ten major milestones (Curran, 1978, p. 135), and promises to provide detailed design specifications for voice interactive systems that maximally contribute to the most appropriate airborne tasks. The voice system and simulator capability will also be suitable for carefully evaluating commercial or contractor-developed voice systems, to see if they adhere to system specifications.

We may summarize this section by noting that, while the airborne applications of voice technology introduce specific advantages and disadvantages (or difficulties), the previous and current projects have laid valuable groundwork and have begun to answer critical questions regarding the tasks and conditions for effective airborne use of voice input (and output) capabilities. The listing and evaluating of critical issues in airborne applications of voice input technology, as outlined in section 4, should aid in the processes of specifying suitable systems, selecting alternative tasks, and evaluating the effectiveness of voice interactions in specific airborne situations.

4. CRITICAL ISSUES, THEIR SEVERITIES, AND LIKELIHOODS OF RESOLUTION

There are many issues and problems raised by airborne applications of speech recognizers. We have seen that certain of those issues have already been addressed in some preliminary studies, and current work is continuing the advancement on some aspects of airborne recognition systems. Yet, much still remains to be done. In this section, we consider a framework for assessing the various issues (subsection 4.1), a specific assessment of issues involved in various aspects of recognition system design and evaluation (subsection 4.2 to 4.8), and an assessment of the criticality and priority of the various issues, and the likelihoods of resolutions occurring in the 1980-1985 time frame (section 4.9).

4.1 A Framework for Assessing Issues

A mere listing of the many problems and issues involved in airborne use of speech recognizers can appear overwhelming and not very informative. Figure 8 provides an organizational framework for categorizing and relating the various problems. Each block in Figure 8 provides a general topic covering a number of issues related to a particular stage in the process of converting task-dictated messages into spoken inputs which, despite interference, must be accurately recognized and used to produce correct machine responses.

The tasks to be handled by spoken commands must be selected to be realistic and useful. Airborne applications impose specific conditions on tasks. Human factors studies are needed to determine what demands can be placed on the human communicator, and how he or she can be most effective in communicating while engaged in the primary tasks of the crew station. Perhaps one of the most critical factors determining the success of a recognition facility is the language used for communicating with the machine. If the language is carefully designed, the complexity of the interactive task can be minimized, and recognition errors can be avoided. In airborne applications, every appropriate constraint should be placed on the recognition task, since the channel limitations and environmental conditions of noise, vibration, g-forces, face masks, and limited-quality audio systems make accurate recognition more difficult to attain. While, as was described in section 2, there

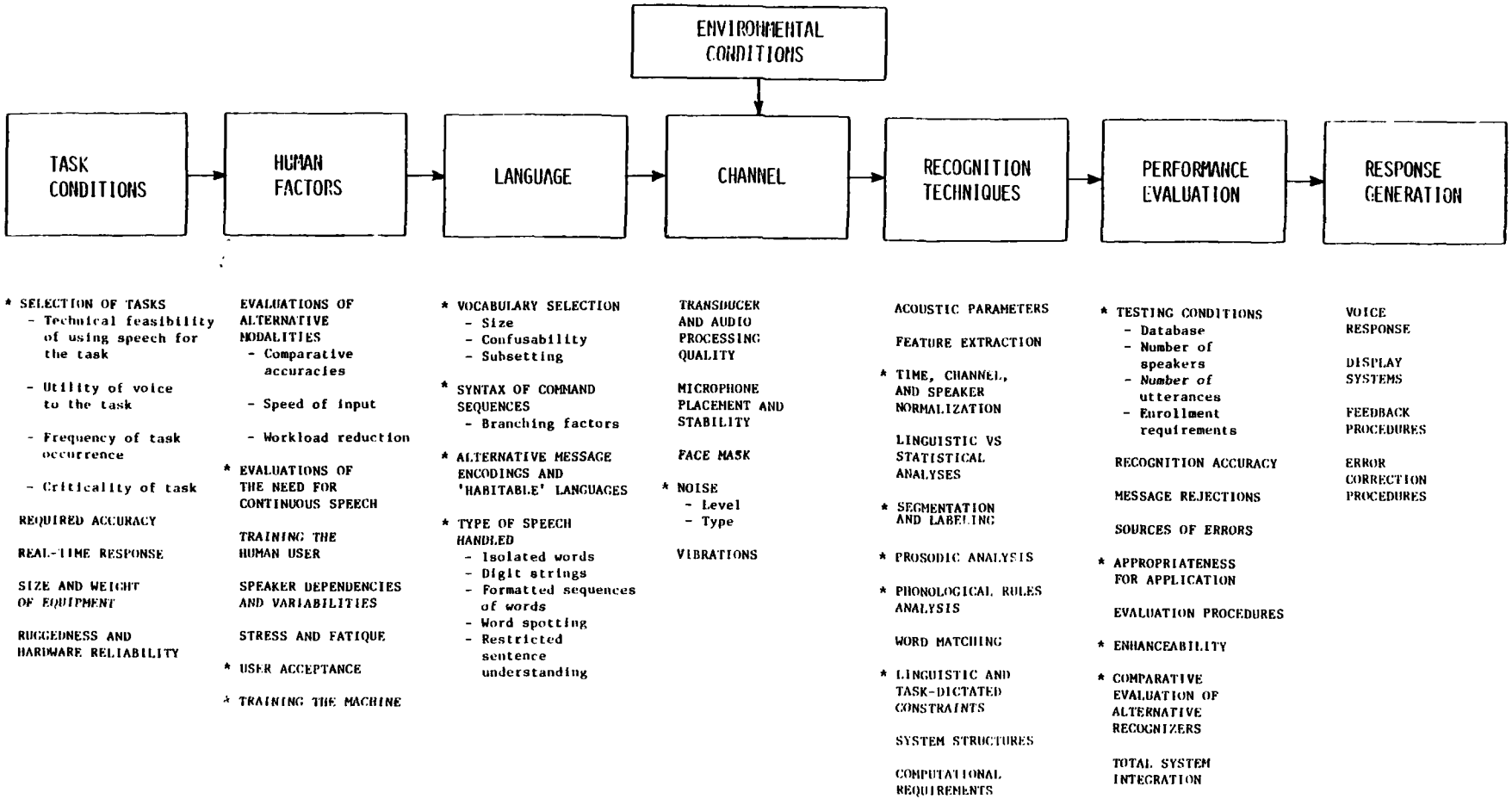


Figure 8. An organizational framework for summarizing the various issues involved in airborne applications of speech recognizers.

are quite accurate recognition techniques for isolated word recognition, even they need substantial improvement and testing to handle airborne applications and time, channel, and speaker variations. Also, for the more versatile facilities for recognizing digit strings, formatted word sequences, or other forms of connected speech, extensive work is needed on all aspects of recognition, especially the "front end" aspects of acoustic phonetic, prosodic, and phonological analysis (or equivalent pattern-matching processes). Performance evaluation is another critical area of recognition system application, and must not only include adequate evaluation of the accuracy of recognition, but also determination of sources of errors and comparative assessment of alternative recognizers. Finally, some consideration should be given to response generation, such as the use of voice response and display systems.

In almost all of these major stages in the human-to-machine communication process, there is at least one, and usually several, key problems to address in programs for airborne application of speech recognition. We shall next consider specific issues and their relative severities, and consider the relative potentials for resolving these issues.

4.2 Task Conditions

The selection of appropriate crew station tasks for use of voice input is one of the most important problems in the application of speech recognizers to airborne systems. The Boeing/Logicon efforts for NADC defined some useful procedures for selecting appropriate tasks, but these (and other) procedures must be applied to the careful selection of high-priority tasks within actual operational aircraft that will need such facilities in the 1980-1985 time frame. The P-3C may or may not be among the finally selected aircraft for introduction of speech recognition facilities, but it is certain that other military airplanes and helicopters will also warrant careful study.

Selection of tasks must be based on the technical feasibility of having speech recognizers handle each task, the utility or improved mission performance produced by use of voice input, the frequency of occurrence of the task during actual missions, the criticality of the task (i.e., will a recognition error be hazardous, or introduce irretrievable system errors), and whether the required accuracy is in line with current or projected recognition

accuracies of commercial recognizers. Technical feasibility involves a judicious match between the needed versatility of the recognizer and the projected capabilities of practical recognizers, on such dimensions as type of speech (isolated words or continuous speech), number and type of speakers, size and confusability of the vocabulary, and other dimensions listed in Figure 8. The Boeing/Logicon procedures for combining these factors into an overall assessment variable may prove useful in such studies.

I would caution program developers against confining projected applications of speech recognition to only those tasks that are amenable to current (1979) technology in isolated word recognition. The technology seems to be on the initial phases of a rapid and substantial upswing, which is reflected in the recent immergence of one commercial recognizer for limited continuous speech and the active efforts at several manufacturers to provide competitive devices for recognizing digit strings and strictly formatted (but continuously spoken) word sequences. The successful demonstration of limited sentence understanding systems in the ARPA project and subsequent research also suggests the potential of more versatile recognizers being possible and commercially available in the 1980-1985 time frame. Other recent work suggests the potential of readily handling many speakers in a speaker-independent manner (i.e., with little or no training), accurately recovering the message on noisy and distorted channels, and handling effects of context and changes in rate of speaking.

Other topics that deserve attention as one adapts recognition capabilities to the specific airborne task include the need for rapid (real time) response, the need for very small size and lightweight equipment, and the demand for ruggedness and high reliability of hardware. Real time response is either already attained or clearly possible, with almost any previously-developed recognizer. This must be demanded in most airborne applications, but it should be no major problem to accomplish. Small size and light weight, rugged hardware should be possible to attain with the current and forthcoming microminiaturized hardware.

While cost has not been a primary factor in most military explorations of speech recognizers, it is clear that, in the past, speech recognizers have usually involved higher initial expense than alternative modalities such as keyboards. Now it appears possible to reduce unit costs on recognizers by

an order of magnitude, so they can be expected to be cost competitive in the 1980-1985 time frame. This may require a substantial ($200K or more) initial expenditure, to fabricate the necessary large-scale-integration hardware. It is encouraging that recent systems for connected speech recognition have had initial costs comparable to (not higher than) initial costs of isolated word recognizers. Active industrial competition should help keep costs down, thus making speech recognition facilities more attractive. The costs of speech recognizers will undoubtedly be a very small fraction of the costs of airborne command and control systems, and need not be considered a primary issue in future program plans.

Thus, the most critical task-related issue in airborne application of speech recognizers will be the careful selection of crew station tasks for which speech input is technically feasible, useful in improving the effectiveness of the crew member, and practical (i.e., required accuracy and other conditions are not beyond the capabilities of projected technology).

4.3 Human Factors Problems

Understanding the human capabilities and limitations is an important aspect of the development of adequate voice interaction systems. Several questions are involved:

- How does voice compare with alternative input modalities, in terms of accuracy, speed of input, and ability to free the human from excessive work loads?
- Is there any real need for continuous speech for the task?
- How much time and effort does the user have to invest in training the machine to recognize his or her voice?
- Does the speaker have to learn to speak in a specific way to have the machine accurately understand what was said? How much adaptation is required of the speaker?
- What effects are found from differences between speakers, and from the variability of any one speaker from time to time?
- What effects do stress, mission duration, and fatigue have on the voice and recognition accuracy?
- What determines user acceptance of a speech recognizer?

As was noted in section 2 and 3, several previous studies have shown the relative advantages and disadvantages of voice input, keyboard data entry,

graphical pen input, and other modalities. Studies with humans interacting with humans have clearly shown that voice input can cut in half the time to interactively perform complex tasks (Chapanis, 1975; Ochsman and Chapanis, 1974). This demonstrates the theoretical ideal that fully versatile speech recognition facilities might aspire to. However, of more immediate interest are the studies (e.g., Welch, 1977) that show how available devices compare as input modalities; such studies have all shown that limited (isolated word) voice input is particularly suitable for more complex tasks such as input of alphanumeric strings, or data input while the human is simultaneously engaged in another primary task which occupies the hands or eyes. These previous studies with current technology show that there is basically no reason not to include voice input, unless the task is so easy that trained users can more rapidly and accurately enter the data on simple keyboards. The limitation on such practical voice input is primarily the error rate of the available isolated word recognizers (which we might reasonably assume is possible to improve). The previous studies have repeatedly shown that flight vocabularies and airborne tasks are amenable to use of current or projected voice input technology.

Currently, NASA Ames is conducting a study of voice input and other modalities under airborne conditions. So is NADC. RADC is also interested. The Royal Aircraft Establishment in England has explored the question, and has plans for further work (Peckham, 1979). So have the ISPENA consortium in France, and other European groups (Haton, 1979). The NATO Working Party AC/243 (Panel III) RSG 10 on Speech Processing has been addressing the issue for years, with several active studies being reported to that group (cf. Peckham, 1979). I predict that all these studies will confirm the view that voice input compares quite favorably with other input modalities, except that: (a) for very simple (e.g., digit entry) tasks, keyboard entry is faster (and currently more accurate) and (b) the utility of voice entry is limited by the error rate of the recognizer, and, to some lesser degree, by the speaking errors introduced by restriction to speaker-dependent isolated word recognition. There seems to be no reason to add to these current modality studies with any further work, unless one can show that a new task being considered for the data input is in major ways not comparable to any of these previously considered tasks. That hardly seems likely for airborne tasks. Voice compares quite favorably with alternative input modalities, in accuracy, speed, and, especially, ability to free the human from excessive workloads. Thus, further modality studies should not have high priority in future work.

A much more important question concerns whether there is a real need for continuous speech input in airborne tasks. This is a question of general concern to all the field of speech recognition. As Beek, and his colleagues (1977) have noted, there is no firmly established operational military need for continuous speech input. Yet, military "needs" are not usually expressed as absolute pre-defined requirements for new technology; rather, they usually are the careful extension of previous experience with limited forms of a similar technology. Continuous speech input is a major step away from simplistic discrete commands by keyboard or thumbwheels. "Needs" might not arise to obvious appearance as independent demands, when the total structure of command and control communications would be altered by voice entry of data. Thus, discrete single words match the established methods, and continuous speech, despite its advantages, might not even be conceived of in some appropriate applications simply because of the major change in interactive philosophy that would be required.

Continuous speech does have some specific advantages that should increasingly be attractive for airborne applications. It is fast, natural, less restrictive, and in consonance with the grouping of discrete commands or words into units composed of multiple words (e.g., the groupings of digits in telephone numbers, in credit card numbers, etc.). Which airborne tasks involve such natural groupings, or require the speed and naturalness-to-the-human of continuous speech? This question is critical, and deserves early attention. It can be answered in part by the careful investigation of the needs of specific airborne tasks, as outlined in section 4.2. However, a limited amount of experimentation is also needed to determine how much benefit comes from using continuous rather than discrete speech, and how versatile the form of continuous speech has to be. Such information about the value of continuous speech is essential if the systems, used in the 1980-1985 time frame (and beyond) is to be prevented from being dead-ended and limited by 1979 technology. Two projects are suggested in section 5 for dealing with this crucial question.

Other human factors issues deserve some attention in future programs. Two such questions concern how much training the machine and the human require before the machine accurately handles a new voice. While speaker independence is very valuable in commercial applications like telephone banking or in military applications such as securing access to secret areas, it is not so crucial in airborne applications. A small number of crew members are involved, and each must be well trained. Both the human and the recognizer can be simultaneous trained in accurate spoken protocol at the same time the human is learning how

to use the airborne systems. No arbitrary new talker can walk up to the machine and attempt to instruct it. Thus, handling speaker differences may be a desireable but non-critical aspect of airborne applications.

More important for airborne applications are the related questions regarding a speaker's variability from time to time or from one set of speaking conditions to another. Commander Wherry of NADC found that the voice degrades after one half hour of airborne conditions, and this reduces recognition accuracy. Despite a large literature regarding the effects of various forms of stress and fatigue on human physiology, including the vocal system, it appears we still do not know exactly what voice changes occur or how those changes might affect speech recognition systems. Currently, a study of stress effects on the voice is being conducted, sponsored by Wright Patterson Air Force Base. Some commercial devices purport to detect psychological stress from characteristics of the voice, but their credibility is seriously in question. Further work on external and internal influences on the crew member's voice seems to be necessary. Studies could couple together tests of stress, fatigue, and normal time-to-time variability, with environmental effects such as noise, g-forces, vibration, and face masks (cf. section 4.5).

Finally, some consideration must be given to overall user acceptance of voice input technology, to assure that developed speech recognition systems will be effectively used. Such user acceptance studies could accompany performance evaluations of recognizers (cf. section 4.7).

4.4 Language Issues

The primary way in which the designer of a voice input facility can ease the difficulty of recognition is by constraining the language of alternative utterances. To a large degree, the task conditions, environment and channel effects, speaker variabilities, and general recognition techniques are beyond the control of a program director or military user, but controlling what utterances are spoken can markedly affect recognition performance. Thus, language issues are among the most critical aspects of a total program plan for developing useful airborne speech recognition facilities.

For each airborne task, a vocabulary must be selected which minimizes the difficulty of recognition, yet maximizes the versatility of communication. Recognition difficulty increases with vocabulary size (how many words must be distinguished), and confusability of words in the vocabulary. The vocabulary can be subsetted, by restricting the number of alternative words that can be said at each point in a discourse. A syntax of allowable command sequences

with a small average "branching factor" can dictate a small "menu" of allowable words at each point in the interactive discourse. There is a definite need for improved methods for measuring (and minimizing) the overall complexity of a recognition task.

Another important language design issue concerns the allowance of alternative ways of saying any intended message, so that the user is not severely constrained in the natural speaking of a message (cf. Wherry, 1975). In it's most general form, this issue involves the habitability, or ease with which a user can learn, and adhere to the constraints of, an interactive language (Watts, 1967). This is a research issue of some long-term interest, but it need not be considered critical in the limited airborne applications of the next few years.

A primary language issue of direct concern in airborne applications is the type of speech to be permitted in the human-to-computer communication. As mentioned while considering human factors in section 4.3, some speech input tasks may require more than the simple small-vocabulary isolated word recognizers available today. Sequences of isolated words or phrases might be composed according to strict syntactic structure rules. Digit strings and strictly formatted word sequences of highly restricted form may be particularly appropriate in some tasks. It is possible that spotting a command word in conversational context might be useful for some tasks. Restricted forms of sentence understanding might also be needed for some complex airborne operations. Much is yet to be resolved regarding technical needs and practical feasibilities of recognition systems over this spectrum of alternative capabilities.

In general, the language design for effective interaction with airborne systems will require close attention. Some valuable projects are suggested in section 5, and these should be considered of major importance. The potential for resolving language issues now seems good.

4.5 Channel and Environmental Conditions

Most speech recognition work has been done under near-ideal conditions of low noise, wide frequency bandwidth, no channel distortions, and high-fidelity electronic equipment. The airborne environment, however, requires recognizers to work in the presence of high noise (around 100 dB or more), poor quality transducers and audio equipment, face-mask-mounted microphones that slip, vibrations, g-forces, and other constraints. In my estimation, previous studies have suggested that vibrations and g-forces are not critical problems; if

present in small amounts, they do not significantly affect recognition, and if present in large amounts, speech input would not (or should not) be attempted. This does mean that during some critical phases of missions when other modalities are also quite difficult to use, voice would not be a potential alternative. Yet, as suggested in the Boeing/Logicon study, the current error rates of speech recognizers prompt caution about their use in critical tasks or situations where error rate might be unusually high. A reasonable cautious position for the early stages of introducing speech recognition into the airborne situations would be to restrict voice input to the more reliable conditions of little vibration and low g-forces. A bolder position which attempted to introduce speech commands during high vibration and g-force would require further simulator studies of how to assure reliable recognition under such challenging conditions. This would seem to be more appropriate after voice has shown its utility in more normal, stable airborne conditions.

A primary reason for variations in recognizer performance with g-forces and vibration has been asserted to be slippage of the face mask (and microphone). Development of a more stable face mask and microphone placement (or, equivalently, procedures for making the recognizer insensitive to such variations) would appear to be an important, but resolvable, aspect of a program for airborne use of voice input.

Noise is one of the most critical environmental problems aboard aircraft, and the value of close-talking, noise-cancelling microphones has been apparent. Current studies are being conducted, at NASA Ames, in France, in England, and in other NATO countries, on the effects of noise on speech recognition. Further work still seems appropriate, given the complexity of the problem, but care must be taken not to redundantly duplicate other current studies. Airborne noise is not simple white noise, so studies such as NASA Ames is undertaking are considering various levels and types of noise, such as helicopter or specific airplane noises.

Noise and other channel distortions cause problems to the simplest acoustic pattern matching methods of utterance identification, and encourage the use of more robust linguistically-based feature extraction schemes in speech recognizers.

4.6 Recognition Techniques

Most aspects of recognition algorithms are in need of improvements (cf. Lea and Shoup, 1979). In general, the acoustic parameters currently monitored in various recognizers seem adequate, and some alternative sets of parameters have been shown to produce about equivalent results in recognition accuracy (Gold-

berg, 1975; White and Neely, 1975). The most accurate commercial recognizers currently use acoustic pattern matching techniques to compare such extracted parametric data with stored templates, or, in the case of the Threshold Technology systems, they extract some linguistically-based features which are then compared with templates obtained from training data. Such acoustically-based recognition techniques are highly susceptible to acoustic changes due to noise, speaker variabilities, rate of speech, microphone movement, channel distortions, and variations introduced by vibration and g-forces. Major issues for improving such schemes include the need to develop procedures for normalizing for time, channel, and speaker variations. Studies of dynamic programming and other normalization procedures should be done on commercial isolated word recognizers and other recognizers of connected speech, with testing done in the presence of noise, speaker variabilities, and airborne channel conditions. It is likely that modest experiments will show ways of significantly improving available algorithms, but that only part of the effects of the various airborne conditions can be compensated for, so that practical airborne operation with current recognition techniques will continue to have higher error rates than under near-ideal laboratory conditions. This warrants developmental studies, and the consideration of alternative techniques that might be less susceptible to acoustic variabilities.

An alternative to strict acoustic pattern matching schemes is found in linguistically-based algorithms. Key information-carrying linguistic features or phonetic-category cues are extracted from the redundant speech signal, so that message-distinguishing aspects of the acoustic data are highlighted while irrelevant aspects are de-emphasized. Phonetic segmentation and labeling can be attempted, to detect message-distinguishing units like vowels and consonants in the speech. In the citation forms of words produced in isolation, most vowels and consonants will be carefully articulated, but even in recognizing isolated words some variation in pronunciations will have to be handled. Also, coarticulatory movements from the center of one phonetic unit to the center of the next unit must be properly characterized to recognize the phonetic structure of a word. Phonological rules can be incorporated into recognizers to account for these variabilities in pronunciation. Prosodic structures such as stress patterns and syllabic divisions of a word can also be useful (Lea, 1979d). Word matching can then be based on comparing the phonetic sequence and prosodic and phonological form of an incoming utterance with the expected pronunciation (or dictionary "transcription") for words in the lexicon of possible words.

All of these acoustic phonetic, prosodic, and phonological aspects of a linguistically-based speech recognizer are considered by most experts to be the top priority topics for further research and development in the next few years (cf. Lea and Shoup, 1979, Chapter 4 and Appendix B). The "front end" of a recognizer (from acoustics to identified words) is thus the primary area for extensive further work. However, unless projects regarding these topics are carefully limited to small tasks that are not substantially beyond the currently-demonstrated capabilities (such as HARPY's limited sentence understanding tasks), this research and development work will have little or no impact on the airborne applications in the 1980-1985 time frame. Thus, isolated words, digit strings, formatted word sequences, and highly restricted sentence understanding tasks are suitable for consideration in a program for initial airborne applications. Higher-level linguistic processing (syntax, semantics, and pragmatics) need not be given extensive additional work, except as how those linguistic and task-dictated constraints can be effectively used to simplify the task of recognition. There likewise seems to be no strong need for extensive work on new or alternative system structures for controlling the various processes involved in linguistically-based recognition.

There seems to be sufficient need for a limited project in continuous speech recognition, such as Boeing/Logicon suggested in their program plan. That project should focus on the development of an appropriate acoustic phonetic, prosodic, and phonological "front end" to a recognizer, coupled with strong linguistic constraints in a finite state or other small grammar. Time, speaker, and channel normalization procedures should be included. The project could attempt to deal with one or more restricted tasks of continuous speech recognition suitable for airborne use, such as recognition of digit strings, formatted word sequences, or sentences from a restricted sentence understanding task. An alternative entirely-acoustically-based template-matching method for connected speech recognition might also be included in such a project.

Thus, critical issues exist in recognition techniques, primarily with regard to the acoustic phonetic, prosodic, and phonological analysis procedures, and the necessary time, speaker, and channel normalization techniques.

4.7 Performance Evaluation

Speech recognizers have usually been assessed on the basis of their recognition accuracy (i.e., percentage of utterances correctly understood) and the percentage of message rejections, for some simple task such as isolated digit recognition. This involves specific testing conditions, such as a selected database of many spoken utterances, uttered by a number of speakers, with some enrollment of the speaker into the system through training utterances. Actually, there is a definite need for the development of standardized testing conditions and databases. Comparative evaluations of alternative recognizers is now a primary topic for further work. Evaluation procedures need to be more fully specified, including how to decide the sources of errors in systems and how to assess the appropriateness of a system for any specific application. Part of the total assessment of a system's adequacy must consider the enhanceability of a specific system or recognition technique, so one avoids developing recognizers that never can readily be applied to a new or bigger task. Finally, total evaluation of a recognizer must give consideration to the way that device can be integrated into the total airborne command and control system.

Besides specifically evaluating the performance of any recognizers developed during projects of the program, the program plan to be outlined in section 5 includes specific projects for (a) comparatively evaluating available commercial recognizers, and (b) devising specific evaluation procedures for assessing total performance of future recognizers.

4.8 Response Generation

While it is beyond the primary purposes of this report to discuss voice response, or spoken output from the computer (or any other computer output modality), we may note that issues exist about the machine's generation of responses to spoken inputs. Previous studies have suggested that voice response should not be used to verify the machine's decisions about spoken inputs, or to provide vocal outputs that confuse the speaker as he or she prepares to say voice commands. Yet, in aircraft the alternative of a visual display means that the crew member's overloaded visual channel will be further taxed, and valuable cockpit space will be required. Some study seems to be called for regarding the trade off between alternative input modalities to the human. Also, specific consideration should be given to the best feedback procedures, so the human can know the machine's decisions about spoken inputs. A related problem is how to devise effective error

correcting procedures, so that a speech recognition error can be corrected. This includes the question of whether a whole connected word sequence or sentence has to be repeated when one word in the utterance is misunderstood. Proper feedback and error correction procedures can make the difference between a "raw" recognition error rate of, say, 2%, and a near-zero operational error rate after verification and correction. These issues might reasonably be included in the performance evaluation efforts of a total program plan in airborne application of recognizers.

4.9 An Assessment of Issues and Potential Resolutions

Figure 9 summarizes the various issues discussed throughout section 4, and offers an overall estimate of the relative significance, and consequent priority, to be attached to each issue. Among the highest priority issues are: the selection of airborne tasks for incorporating voice input; type of speech handled; confusability of words; noise effects; and advanced techniques for acoustic phonetic analysis. Closely following those issues are ones concerned with stress and fatigue effects; evaluation of the need for continuous speech; prosodic analysis; phonological analysis; normalization procedures; conditions for testing systems for performance evaluation; and comparative evaluations of alternative recognizers. Other issues have lesser priorities. Within each major topic area of Figure 9, the issues are ordered in approximate order of decreasing significance.

No conceptual bottlenecks or major long-term difficulties of resolving the various issues are apparent, except perhaps for the difficulties and highly intense effort involved in acoustic phonetic, prosodic, and phonological analysis procedures. Our next task is to outline a program plan that can deal with all these issues or problems involved in developing an adequate technology of voice input in airborne environments.

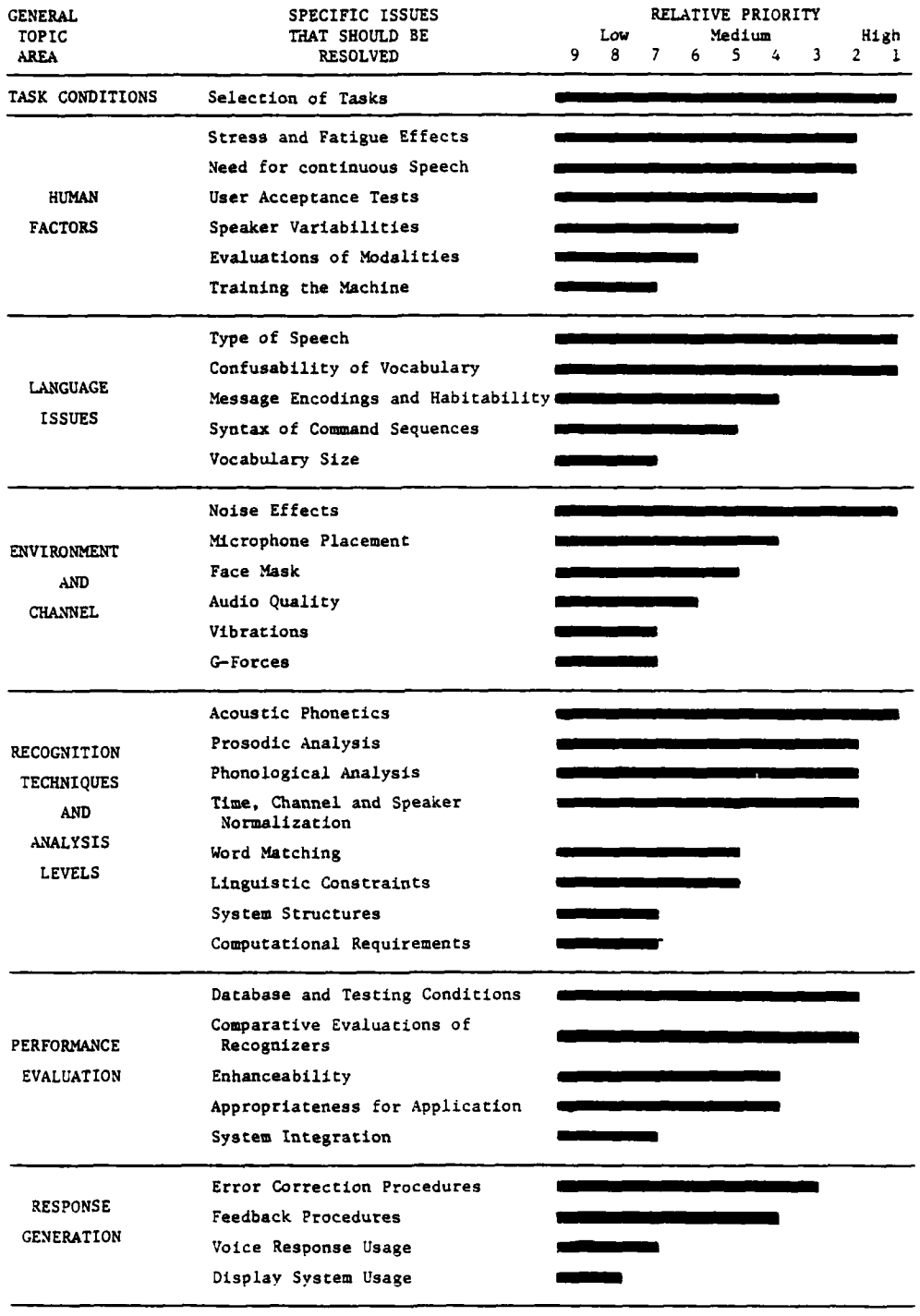


Figure 9. Primary Issues in Airborne Applications of Speech Recognition, with Relative Priorities Indicated by Lengths of the Bars.

5. SPECIFIC RECOMMENDATIONS FOR AIRBORNE APPLICATIONS

Based on the discussion of issues as presented in section 4, we can define an overall program for developing a technology for airborne application of speech recognition. The nature of important additional work can be outlined, and some estimate given of the magnitude of required effort to resolve the critical issues. No claim can be made that such a program will be fully comprehensive, given the methodology used in the present study, but the program plan can help identify important problems, and define projects that should resolve those problems. In section 5.1, a program plan is offered, including a list of projects and some guidelines about time schedule and level of required effort. Each project is briefly described in section 5.2, and a summary and prospectus is given in section 5.3.

5.1 A Program Plan for Developing Airborne Voice Input Systems

Figure 10 lists a number of projects needed to achieve effective airborne applications of speech recognizers. The scheduling of the various projects is suggested, and the approximate level of effort is indicated, along with the roughly-estimated costs (based on $100K covering the cost of one man-year effort by senior personnel, plus associated junior personnel, overhead, computer and hardware costs, and other expenses). Fiscal year 1979 is indicated in the figure to reflect some efforts that have already begun or that have been proposed. The program is scheduled with an approximately equal work load in each year, totalling around $1.5 million per year, at near minimum, to approximately $2.5 million per year.

The listing of projects in Figure 10 is based on the priorities of issues previously outlined in Figure 9. Highest priority tasks are shown by bold black bars, medium priority by crosshatched bars, and low priority ones by white bars. Some critical issues do not require extensive workload despite their importance, and some lower-priority issues, by the very nature of the work involved, do require a fairly heavy committment of funds. All of the projects shown in Figure 10 have some substantial significance to an overall program in the development and application of airborne speech recognizers. Additional lower-priority work could be possible.

The five year effort presented in Figure 10 should provide necessary technology, feasibility demonstrations, initial system testing in (simulated) airborne environments, user studies, and preliminary performance evaluations. Operational deployment and extended use is not reflected in the plan, but

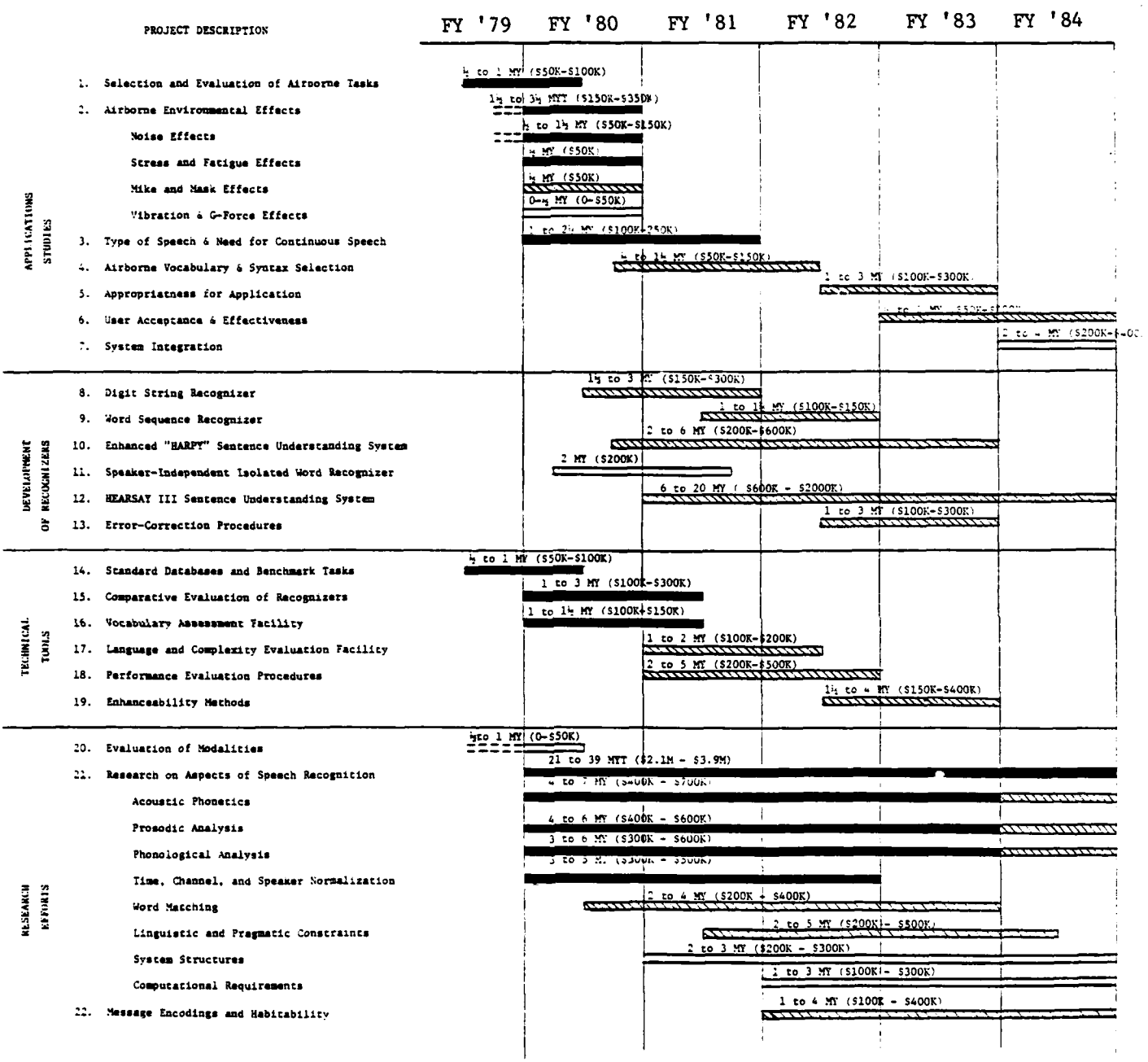


Figure 10. A Program Plan for Achieving Effective Airborne Applications of Speech Recognizers. High priority projects are shown by solid black bars; medium priority by crosshatched bars, and low priority by open (white) bars. Levels of primary technical effort are shown in man years (MY), and costs in thousands of dollars ($1K) are estimated, based on $100K ($100,000) per man year (junior personnel staffing and other costs are also included in the $100K/MY figure).

would presumably begin near the end of the 1980-1985 time frame.

This program plan assumes a reasonable level of cooperation among various agencies involved in airborne applications, so that redundant projects are avoided. There is no assumption that all of the indicated projects must be done at, or sponsored by, NADC. Indeed, the brief descriptions of each project, to be given in section 5.2, will probably naturally suggest alternative agencies, industrial groups, and academic or research-oriented groups that might be appropriate, based on previous interests and efforts at the various organizations.

5.2 Descriptions of Recommended Projects

In this section, we shall briefly describe the specific projects listed in Figure 10, and suggest general project justifications and possible methods of attack on the problems. Notice that the projects in Figure 10 are divided into ones dealing with: applications studies (sections 5.2.1 to 5.2.7); development of recognizers (sections 5.2.8 to 5.2.13); technical tools to help design and evaluate recognizers (sections 5.2.14 to 5.2.19); and research efforts (sections 5.2.20 to 5.2.22).

5.2.1 Selection and Evaluation of Airborne Tasks - Much of the recommended work depends upon the early selection and analysis of airborne tasks that appear most promising for application of voice input technology. Previous work such as the Boeing/Logicon study has suggested the methodology for such a study, but the recommended project shown in Figure 10 refers to careful study of actual candidate aircraft and the most promising crew station tasks. Currently, work is under way at NASA Ames along these lines, as well as within the NADC VIST program. Presumably, continued work would be appropriate. Some of the constructive discussion of the Boeing/Logicon methodology and conclusions, as presented in section 2 of this report, might help guide this applications study. A level of effort comparable to the earlier Boeing/Logicon study seems appropriate (i.e., about one-half to one man-year; a man-year is estimated to cost around $100,000, abbreviated $100K).

5.2.2 Airborne Environmental Effects - Previous studies have acknowledged the need for early attention to the effects of airborne environmental conditions on the effectiveness of speech recognizers. Further simulation studies seem to be needed, and NASA Ames, NADC, WPAFB, and other groups have current efforts on these problems. The program plan of Figure 10 indicates the need

for early attention to continued work on these environmental and channel issues. A single project is suggested, with concentration on the high priority topics of noise and stress and fatigue effects, but some attention to mask and microphone positioning. (Current work is already being pursued on stress and tatigue factors, through WPAFB.) Vibration and g-forces are still of concern, but, as argued in section 4, early work on application of voice technology might avoid use of voice input for critical tasks performed under such conditions. It is estimated that, depending on the applications selected and the predicted airborne environmental conditions when voice input is used, this project might involve 1½ to 3½ man years of total effort (MYT), divided among tasks as shown in Figure 10. Part of this effort may be accomplished through the current NASA Ames and NADC studies, either at the agencies involved (using appropriate cockpit simulators) or at contracted groups with experience in operational airborne systems and voice input technology.

5.2.3 Type of Speech and Need for Continuous Speech - It is crucial to the development of effective voice command and control systems that we first define the type of speech to be handled. Different approaches and demands will be involved for: isolated words taken from a small vocabulary, or from a large vocabulary, or from a subsetted vocabulary used with a complex command syntax; connected digit strings; strictly formatted sequences of words spoken continuously; spotting of key command words in connected speech; and continuously spoken English sentences. While we may anticipate that practical work in the 1980-1985 time frame will be dominated by use of the simplest possible types of speech (i.e., isolated words and digit strings), the recommended project on types of speech and the evaluation of the need for continuous speech should be important in defining the most suitable forms of recognizers. Such a project should include study of the airborne tasks that seem most in need of connected speech, to see how important the connected flow of commands should be. This could be done either by NADC or another government agency, or by a contractor that is not already biased strongly either towards current IWR approaches or towards research on fully versatile speech. Another major aspect of this study should involve experimental studies of the effectiveness of restricted human-to-human or human-to-machine interactions, under varying conditions such as increasing sizes of vocabularies, simple versus complex word sequences or sentences, and various grammatical flexibilities. It is crucial to know how much language and interactive ability is really needed for

effective voice command of airborne machines. Alphonse Chapanis at Johns Hopkins University has been working for years on the question of restricted communication with machines, including his recently showing that humans communicated as well when restricted to 300 word vocabularies as when permitted an unrestricted vocabulary. (Chapanis, et al., 1977). The tasks on which further interaction experiments are to be conducted should be required to be relevant airborne tasks. It is estimated that this total project might require 1 to 2½ man years.

5.2.4 Airborne Vocabulary and Syntax Selection - Another application study of some interest concerns the selection of a vocabulary and command syntax for use in the most promising airborne applications. Vocabulary design is a fairly straightforward process once the task and the full protocols of crew station interaction are known. Subsetting of the full vocabulary into menus of alternative next words can be done for each stage in the interactive discourse (i.e., each node in the syntax tree). However, for accurate recognition and maximum success in voice control procedures, the vocabulary and subvocabularies must be designed to minimize the likelihood of similar words being confused. Vocabulary sizes and confusabilities can be systematically designed using the vocabulary assessment facility to be described in section 5.2.16, so this project has been delayed until that project (number 16 in Figure 10) is at an advanced stage. This vocabulary and syntax selection has also been timed to take advantage of the task selection, study of environmental factors, and initial aspects of the study of types of speech (projects 1 to 3 in Figure 10). Depending in part on the number of airborne tasks and the size of the vocabularies and interactive language, and in part on the availability of the vocabulary assessment facility, this project might require ½ to 1½ man years. With the vocabulary assessment capability to be recommended in project 16, this project might be rapidly and inexpensively accomplished at NADC or wherever that capability is operational.

5.2.5 Appropriateness for Application - In conjunction with the performance evaluation of sophisticated recognizers that are developed during the project (e.g., project 18 in section 5.2.18), it seems important to conduct an applications study to determine whether a recognizer (or alternative recognizers) will perform well and truly be appropriate to the planned airborne applications. Project 5 is intended to test the appropriateness of recognizers for practical applications, by careful study under accurately simulated or fully operational

airborne situations. As the first stage of specific technology transfer, such a study would involve ruggedized equipment of small size and weight used under realistic conditions. Depending in part on how many recognizers must be tested and what special purpose design is involved, this applications study might require 1 to 3 man years. It might be done either at a governmental agency with a centrifuge and total simulator facility, such as NADC, or at an industrial firm with operational experience. It obviously must be delayed until the program has produced near-final versions of the recognizers being designed.

5.2.6 User Acceptance and Effectiveness - Closely allied with the previous project, which will evaluate the machine working in operational or near-operational conditions, the user's response and effectiveness with speech recognizers should be studied for airborne applications. No system can ever be useful in real military operations if the user will not willingly use it. Pilots and crew members are known to be justly sceptical of new unproven technology. Also, experience shows that devices which work well in the laboratory, with cooperative trained users, may degrade in performance when operated by the ultimate users in the practical environment. Depending on the number and nature of systems to be evaluated, this user evaluation study might involve from ½ to 2 man years of effort, preferrably conducted by a governmental agency or an experienced testing laboratory.

5.2.7 System Integration - A final issue in the application of recognition technology is the integration of the recognizer into the total airborne command and control system. Such an integration project is proposed for the final year of the program, to integrate the initially selected form(s) of recognizers into the airborne system. This might involve 2 to 4 man years.

5.2.8 Digit String Recognizer - Besides the obvious purchase of commercially available isolated word recognizers for testing in airborne operational conditions, there will probably be a need to develop specific forms of new, more advanced recognizers. The first-priority advancement would seem to be a digit string recognizer. It is too early to tell whether the recently announced Nippon Electric Company system will meet some or all of the stringent demands of an airborne digit string recognizer. Other companies are currently

working on commercial products of this form, so that at least several initial forms of digit string recognizers should be available within the next two years. Either these straight forward extensions from current IWR technology, or a totally new approach, may provide an accurate prototype of this limited form of connected speech recognizer. A project is recommended to adapt available devices or independently develop a suitable recognizer, to achieve highly accurate digit string recognition under conditions suitable for use in airborne applications. This project might involve 1½ to 3 man years, and might be conducted either by one of the commercial manufacturers of recognizers, or by a product-oriented development group at a system-developing industrial firm. Notice that this project, and the other development projects listed in Figure 10, are contingent upon the study of airborne applications needs (project 3) showing that some form(s) of continuous speech should be incorporated into airborne systems. Even if initial airborne uses of speech input do not require connected speech, this development of connected-speech recognizers is well warranted if at some time and for some tasks it is anticipated that connected speech will be useful aboard aircraft.

5.2.9 Word Sequence Recognizer - Closely allied with digit string recognition is the task of recognizing strictly formatted connected word sequences, such as alphanumeric strings, commands with few alternative words allowed in each position, etc. The more complex vocabulary and the use of syntactic constraints make this type of recognizer somewhat more challenging than digit string recognition, but the techniques can be quite similar. Development of a word-sequence recognizer could even be "piggy-backed" on the project in digit string recognition. Thus, this effort might involve a fairly modest additional effort of only 1 to 1½ man years. For less-constrained forms of sequence recognition, additional research and development might be required.

5.2.10 Enhanced "HARPY" Sentence Understanding System - At the forefront of current demonstrated capabilities in continuous speech recognition is the HARPY speech understanding system developed at Carnegie-Mellon University. The tasks for which HARPY has demonstrated considerable success are perhaps at the extreme high end of the necessary capabilities of immediate airborne interactions, but a HARPY-like system represents the next logical step up from highly constrained digit-string and formatted-word-sequence recognizers. It is worthy of further development and use.

However, as the original developers of HARPY have astutely observed (Lowerre and Reddy, 1979), HARPY could profit from some significant improvements. It is currently being re-designed at CMU, for near-real-time use on a minicomputer. Vital refinements ("incremental compilation procedures") are needed so that the whole pronunciation network need not be re-compiled each time a single word, or structure, or pronunciation is altered. Automatic acquisition of network knowledge has also been called for. The acoustic phonetic analysis needs substantial improvement, and some promising refinements in sub-phonetic templates, phonological rules, and search procedures are possible. This author has also devised some methods for further use of prosodic information in HARPY. A low-key effort is underway at SCRL and the University of Southern California (USC) to re-implement HARPY on the SCRL PDP-11/45 system, with possible enhancements to be developed as a part of speech recognition coursework at USC. CMU is also continuing work on HARPY.

A project on enhancement of a HARPY-like system, with airborne applications in mind, seems to be very appropriate for the recommended program. Depending on how much enhancement of HARPY is attempted, and how much work is devoted to developing a form of HARPY suitable for specific airborne tasks, this promising project might involve 2 to 6 man years of effort. The schedule in Figure 10 shows the project ending a year before the end of the program, to permit transferring this advanced technology into initial (or later) recognizers to be tested in airborne applications. However, continued work on this type of advanced system could follow, as preparations are made for more sophisticated ("second-generation") airborne voice input facilities.

5.2.11 Speaker-Independent Isolated Word Recognizer

While speaker independence does not appear to be a high priority topic for airborne applications, it is useful to minimize the time and effort involved in each user having to train a speaker-dependent isolated word recognizer. A 2 man year project is proposed to accomplish such a development. This project could begin early in FY '81, as shown in Figure 10, since it is a limited extension beyond current technology. However, it might also be appropriate to delay this project until the end of FY '81, and take advantage of the results from current commercial efforts in speaker-independent recognition. Discussions with manufacturers of commercial recognizers have indicated that several of them are working on, or interested in offering, speaker-independent recognizers, and their commercial results might be exploited in FY '82. Otherwise, current advanced work at other research and

development groups might be brought to bear on this problem, through a project at one of the manufacturers, or a system development group.

5.2.12 HEARSAY III Sentence Understanding System - In combination with the research on various aspects of speech recognition, to be described in section 5.2.21, an advanced system development project is recommended, based on the best of technology developed during the ARPA SUR project. A "HEARSAY III" restricted sentence understanding system should be developed, based on the basic modular structure and "blackboard" model of system control used in the successful HEARSAY II system developed at Carnegie-Mellon University. Such a system has been argued to be the best for future work on an advanced speech understanding system (Lea and Shoup, 1979); cf. also section 2 of this report), and could also readily incorporate some of the excellent components of the BBN HWIM system. This advanced system would be suitable for further work with second or third generation devices for airborne voice command and control.

This ambitious but promising project in restricted sentence understanding (presumably with versatile facilities for continuous-speech communication with airborne computers) could involve about 6 to 20 man years of effort. Coupled with the research recommended in project 21, this would be similar to the continuous speech recognition project proposed in the Boeing/Logicon study.

5.2.13 Error-Correction Procedures

Every recognizer will make some errors, and a useful area for further development in recognition techniques would be to devise procedures for correcting some or all of those errors. Current speech recognizers use decision feedback to the human for detection of errors. Then, correction involves the human instructing the machine that an error has been made, and speaking the correct word again, until it is properly identified. One concern in error correction is thus the form of recognizer feedback of decisions, either by way of a voice response or a display of the recognition decision. Some work suggests that voice response may be confusing, but the alternative of another display aboard aircraft is not attractive. Also, with continuous speech, a long sequence of words may have to be repeated, unless procedures can be devised for indicateng which word is to be corrected. Feedback procedures need further study.

In addition, it is possible to design command protocols so as to automatically detect some errors. "Check digits" in carefully formatted digit strings are an example of such error detection, and error-correcting procedures. Texas Instruments has used such check digits in their system for digit string recognition and speaker verification. Further work seems warranted on such ideas.

This work on developing error correction procedures is estimated to require 1 to 3 man years of effort.

5.2.14 Standard Databases and Benchmark Tasks - A variety of "technical tools" or service capabilities are needed that are not tied to the development of one specific recognizer. These will help advance the total technology, permitting evaluation of available recognizers and guiding the design and redesign of successively more challenging tasks. One such technical tool that could help advance the total technology is the compilation of "standard" databases and "benchmark tasks" for evaluating alternative recognizers. The need for such tools has been discussed previously (Lea and Shoup, 1978, 1979; also, cf. section 4 of this report). The IEEE has a subcommittee working on appropriate databases. This topic seems to be a top priority concern for the field, and for the recommended program. It would involve only about ½ to 1 man year for selection of initial databases and benchmark tasks. This type of tool could conceivably be developed by the commercial manufacturers, though there is a strong danger of bias under such an approach. Alternatively, a government agency like NADC, or an independent research or development laboratory, could properly undertake this task.

5.2.15 Comparative Evaluation of Recognizers - Recent studies (Lea and Shoup, 1978, 1979) and the current concerns for credibility for commercial manufacturers of recognizers have pointed to the crucial need for methods for comparatively evaluating available recognizers. Using the databases and benchmark tasks, and including airborne application tasks, this high-priority project could establish which recognizers are good for which tasks. One would then not need to rely on manufacturers' claims, but would have independent evidence about performance. Also, effects of tasks of various complexities could be empirically determined. Depending on how thorough and systematic these evaluations are, the various tasks included, and the number

of recognizers considered, this project might take about 1 to 3 man years. It could be done at an independent research or development laboratory, or testing facility, or at a government agency.

5.2.16 Vocabulary Assessment Facility - A novel idea has been proposed to NADC by SCRL, for the development of a method of determining the confusability of words in various vocabularies and predicting the likelihood of recognition errors. The similarity in pronunciations of words is determined from their dictionary forms, and probabilities of confusions are established. Words that are readily confused can then be replaced by other words, and new confusability tests made on the modified vocabulary, until a satisfactory vocabulary results. Such a capability can be used to predict how a recognizer will perform on a new vocabulary, such as a specific flight vocabulary, given knowledge of its performance with a previous vocabulary, such as the ten digits. It also could predict performance under new environmental conditions (or new speakers), given performance under old conditions, plus minimal knowledge about the effects of the environmental changes on confusability of pronunciations.

This vocabulary assessment methodology thus permits systematic assessments of many different vocabularies for various tasks, without the need for the usual expensive experimental processes of: obtaining recordings of several speakers saying all vocabulary words; processing many such utterances through the recognizer; determining recognition accuracy scores; and repeating the process again and again for each new set of conditions (noise, stress, vibration, g-forces, etc.) and each new vocabulary. Thus, instead of unaided designing and extensively testing of each new vocabulary for each new task, a systematic procedure can provide guidelines for assessing many vocabularies.

Initially, this capability for vocabulary assessment could be restricted to work with isolated words, but subsequent refinements could permit use with continuous speech recognizers, for which context and the coarticulatory effects of smooth flow of speech must be accounted for. Also, syntactic trees could be systematically altered to minimize confusabilities of subvocabularies used at various stages in discourse.

This project might require about 1 to 1½ man year of effort, and could be naturally followed by the next project to be described.

5.2.17 Language and Complexity Evaluation Facility - Determining confusability of words is one aspect of assessing the total complexity of a recognition task. As has been discussed previously (Lea and Shoup, 1979; also, section 2 of this report), an important aspect of total evaluation of a recognizer is to determine the complexity of the problem it addresses. Ninety percent accuracy on a simple problem may be less satisfactory than 60% accuracy on a difficult problem. Confusability of words, syntactic branching factors (Goodman, 1976), entropy (Flanagan, et al., 1979), and other measures could help assess alternative interactive languages, and could help predict recognizer performance as tasks change in complexity. Both theoretical and empirical studies are needed on procedures for assessing languages and task complexities. It is estimated that basic work on such a project might involve about 1 to 2 man years.

5.2.18 Performance Evaluation Procedures - Experts agree that an important aspect of future recognition work concerns comprehensive procedures for performance evaluation. Recognition accuracy is not the total picture; sources of error in recognition need to be determined, and the contribution or weakness of each component in the system must be established. Computer scientists have developed a variety of techniques for studying system performance, such as causality analysis, ablation studies, benchmark tasks, null and optimum models, analysis of variance, and operations research models (Newell, 1975). These and other techniques need to be applied to speech recognizers. A project of 2 to 5 man years is proposed for developing advanced performance evaluation procedures, and applying them to evaluate and help improve the recognizers developed in the program.

5.2.19 Enhanceability Methods - Some recognition techniques have been criticized for being "dead-end approaches", while other more complex approaches are justified on the basis of being more open-ended and suitable for use with other, bigger recognition tasks. "Enhanceability" is one term for the capability of a system to be readily adapted to either new tasks of comparable difficulty, or new tasks of higher complexity. Procedures are needed for straightforward extension of limited capabilities to a series of successively more challenging tasks. A project is recommended to determine what is required to accomplish enhanceability, and to apply those techniques to available recognizers. Such a project could be a natural extension of projects 16, 17, and 18, and should require about 1½ to 4 man years.

5.2.20 Evaluation of Modalities - Although the evidence favoring voice input to machines seems to be very substantial, so that evaluations of alternative modalities is not a high priority issue, it is obvious that any additional study that is to be done in this area should be done soon or not at all. A study of about ½ to 1 man year might be required, though the current governmental work probably adequately covers half of that effort, so less than one half man year of further work might be sufficient, in the beginning of the program.

5.2.21 Research on Aspects of Speech Recognition - A major research effort is needed to develop advanced recognition techniques sufficient for handling various forms of continuous speech (especially for sentence understanding). The magnitude of the problem of developing the knowledge and technology needed for advanced continuous speech recognizers is so large that an intensive total five year effort of 21 to 39 man years is recommended. Top priority must be given to acoustic phonetic analysis techniques (4 to 7 man years), prosodic analysis (4 to 6 man years), phonological analysis (3 to 6 man years), and procedures for time, channel, and speaker normalization (3 to 5 man years). Expert opinions about the most promising techniques for such improved "front end" analyses were reviewed by Lea and Shoup (1979). Other moderately important research and development work should be done on improved word matching procedures (2 to 4 man years), and linguistic and pragmatic constraints (2 to 5 man years). Though of lower priority since several adequate methods have been demonstrated, system structures and computational requirements deserve further work (estimated at 2 to 3 and 1 to 3 man years, respectively).

This research effort may have some direct spinoff into improved operational recognizers within the 1980-1985 time frame, but the primary impact of such research would be evident in later airborne speech recognizers.

5.2.22 Message Encodings and Habitability - A final research project of some interest in the recommended program deals with the flexibility in spoken communication that humans take for granted. A particular message may be said in many different ways in human conversations. Machines, on the other hand, traditionally permit only one (or a very few) ways of saying most messages, with strict adherence to highly formatted symbol strings. The work of Robert Wherry at NADC produced a facility that permitted alternative encoding of messages, though the language was necessarily still quite restricted.

Truth tables and semantic associations permitted the machine to comprehend what was meant rather than merely accepting what was explicitly and correctly said within a fixed format. A truly versatile voice input facility will probably require such a flexibility in how to encode a message.

Also, there has long been an acknowledged need for "habitable" languages, which are easy for users to learn and for which it is easy to keep within the syntax and other constraints. A research project on the understanding of habitability and the use of flexible message encoding procedures is suggested. We need to know what makes a restricted language habitable and flexible, and how to use that knowledge in speech recognition facilities. Such a project is estimated to require about 1 to 4 man years.

5.3 Summary and Prospectus

As ambitious as the recommended program is, obviously it is not exhaustive of all possible work in speech recognition. The program hopefully does show the crucial issues, their relative severities and priorities, and how it is possible to resolve those issues. Some of the research work and lower priority work on technical tools is supportive of the longer-range goals of the NADC program (cf. Curran, 1978). However, the applications studies and developments of specific recognizers, and the high priority technical tools (e.g., vocabulary assessment and comparative evaluations of recognizers) promise to have early impact on the mainstream of the NADC work on a hierarchy of demonstrated recognition facilities operating in (simulated) airborne situations.

The projects proposed in this section are based on the structured list of key issues presented in section 4, and represent critical advancements on the best of the previous work described in sections 2 and 3 of this report. Previous work has produced a spectrum of recognition capabilities, ranging from accurate commercially available isolated word recognizers, up to fairly versatile sentence understanding systems developed in the ARPA SUR project. Clearly, the earliest forms of recognizers for airborne applications will involve the simplest, most-well-established devices and feasibility models for isolated word recognition, digit string recognition, recognition of formatted word sequences, and possible limited sentence understanding systems such as HARPY.

While further study is warranted of the effects of airborne conditions on recognizer performance, it seems clear that the airborne applications actually offer some strong potential advantages for voice input, while the apparent

disadvantages are probably amenable to known solutions. We may conclude that airborne applications of speech recognition facilities are possible and well worth pursuing, and that the guidelines given in this report will hopefully help to advance the practical use of limited voice input aboard aircraft. The speech recognition technology has advanced to a level that warrants a vigorous and timely program in the 1980-1985 time frame.

6. REFERENCES

Beek, B., E.P. Neuberg, and D.C. Hodge (1977), An Assessment of the Technology of Automatic Speech Recognition for Military Applications, IEEE Transactions Acoustics, Speech, and Signal Processing, ASSP-25, Number 4, 310-322.

Breaux, R. (1978), Laboratory Demonstration Computer Speech Recognition in Training, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control System Application, (R. Breaux, M. Curran, and E.M. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, December 6-8, 1977.

Bridle, J. and J.B. Peckham (1978), Personal communications.

Chapanis, A. (1975), Interactive Human Communication, Scientific American, Vol. 232, 36-42.

Chapanis, A., R.N. Parrish, R.B. Ochsman, and C.D. Weeks (1977), Studies in Interactive Communication: II. The Effects of Four Communication Modes on the Linguistic Performance of Teams during Cooperative Problem Solving, Human Factors, 19, No. 2, 101-126.

Coler, C.R., E.M. Huff, R.P. Plummer, and M.H. Hitchcock (1978), Automatic Speech Recognition Research at NASA-Ames Research Center, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaux, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 143-170.

Curran, M. (1978), Voice Integrated Systems, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaus, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 123-137.

Davis, K.H., R. Biddulph, and J. Balashek (1952), Automatic Recognition of Spoken Digits, The Journal of the Acoustical Society of America, Vol. 24, 637-645.

Denes, P. and M.V. Mathews (1960), Spoken Digit Recognition Using Time-Frequency Patterns Matching, The Journal of the Acoustical Society of America, Vol. 32, 1450-1455.

Doddington, G. (1976), Personal Identity Verification Using Voice, presented at ELECTRO 76, Boston, Mass.

Dudley, H., and S. Balashek (1958), Automatic Recognition of Phonetic Patterns in Speech, The Journal of the Acoustical Society of America, Vol. 30, 721-739.

Feuge, R.L., and C.W. Geer (1978), Integrated Applications of Automated Speech Technology, Final Report and Program Plan on ONR Contract N00014-77-C-0401, Boeing Aerospace Company (with Logicon, Inc.), Seattle, WA.

Flanagan, J.L., S. Levinson, L.R. Rabiner, and A.E. Rosenberg (1979), Techniques for Expanding the Capabilities of Practical Speech Recognizers, in Trends in Speech Recognition (W.A. Lea, Editor), Englewood Cliffs, NJ: Prentice-Hall, Chapter 18.

Forgie, J.W. and C.D. Forgie (1959), Results Obtained from a Vowel Recognition Computer Program, The Journal of the Acoustical Society of America, Vol. 31, 1480-1489.

Forgie, J.W. and C.D. Forgie (1962), Automatic Method of Plosive Identification, The Journal of the Acoustical Society of America, Vol. 34, 1979 (A).

Goldberg, H.G. (1975), Segmentation and Labeling of Speech: A Comparative Performance Evaluation, Technical Report CMUCSD, Ph.D. Dissertation, Carnegie-Mellon University, Pittsburgh, PA.

Goodman, G., D. Scelza, and B. Beek (1977), An Application of Connected Speech to the Cartography Task, Proceedings of the 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, CT, 811-814.

Grady, M.W., M.B. Hicklin, and J.E. Porter (1978), Practical Applications of Interactive Voice Technologies--Some Accomplishments and Prospects, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaux, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 217-233.

Haton, J.P. (1979), Speech Recognition Work in Western Europe, in Trends in Speech Recognition, (W.A. Lea, Editor), Englewood Cliffs, NJ: Prentice-Hall, Chapter 24.

Hemdal, J.F. and G.W. Hughes (1967), A Feature Based Computer Recognition Program for the Modeling of Vowel Perception, in Models for the Perception of Speech and Visual Form, W. Wathen-Dunn, Editor, Cambridge, MA: M.I.T. Press.

Herscher, M.B. (1978), Real-Time Interactive Speech Technology at Threshold Technology, Inc., Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaux, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 217-233.

Hughes, G.W. (1961), The Recognition of Speech by Machine, Technical Report 395, Research Laboratory of Electronics, M.I.T., Cambridge, MA.

Itakura, F. (1975), Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Transactions on Acoustics Speech, and Signal Processing, Vol. ASSP-23, 67-72.

Lea, W.A. (1979a), Editor: Trends in Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall.

Lea, W.A. (1979b), The Value of Speech Recognition Systems, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 1.

Lea, W.A. (1979c), Speech Recognition: Past, Present, and Future, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Chapter 4.

Lea, W.A. (1979d), Prosodic Aids to Speech Recognition, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 8.

Lea, W.A. and J.E. Shoup (1978), Recommendations for Advancing Speech Recognition, Journal of the Acoustical Society of America, 63: Supplement 1, S78(A).

Lea, W.A. and J.E. Shoup (1979), Review of the ARPA SUR Project and Survey of Current Technology in Speech Understanding, Final Report on ONR Contract

N00014-77-C-0570, Speech Communications Research Laboratory, Los Angeles, CA.

Lesk, M.E. and C.A. McGonegal (1976), User Operated Directory Assistance, Murray Hill, NJ: Bell Laboratories Technical Memorandum.

Levinson, S.E., A.E. Rosenberg, and J.L. Flanagan (1977), Evaluation of a Word Recognition System, Proceedings of the 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, CT.

Lowerre, B.T. and D.R. Reddy (1979), The Harpy Speech Understanding System, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 15.

Martin, T.B. (1976), Practical Applications of Voice Input to Machines, Proceedings of the IEEE, Volume 64, Number 4, 487-501.

Martin, T.B. and E.F. Grunza (1974), Voice Control Demonstration System, Technical Report AFAL-TR-74-174, Wright-Patterson Air Force Base, Ohio.

Martin, T.B. and J. Welch (1979), Practical Speech Recognizers and Some Performance Evaluation Parameters, Trends in Speech Recognition,(W.A. Lea, Editor), Englewood Cliffs, NJ: Prentice-Hall, Chapter 3.

Medress, M.F. (1969), Computer Recognition of Single-Syllable English Words, Ph.D. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology.

Medress, M.F. (1972), A Procedure for Machine Recognition of Speech, Conference Record of the 1972 Conference on Speech Communication and Processing. Newton, MA: IEEE Catalog Number AD-742236, 113-116.

Medress, M.F. (1979), The Sperry Univac System for Continuous Speech Recognition, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 19.

Montague, H. (1977), Voice Control Systems for Airborne Environments Final Report RADC-TR-77-189, Final Report on Contract No. F30602-76-C-0127, by Scope Electronics, Inc., Reston, VA.

Newell, A. (1975), A Tutorial on Speech Understanding Systems, Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium (D.R. Reddy, Editor), New York: Academic Press, 3-54.

Nye, J.M. (1979), The Expanding Market for Commercial Speech Recognizers, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 20.

Ochsman, R.B. and A. Chapanis (1974), The Effects of 10 Communication Modes on the Behaviour of Teams During Co-Operative Problem-Solving, International Journal Man-Machine Studies, Volume 6, 579-619.

Otten, K.W. (1966), Automatic Recognition of Continuous Speech, Technical Report AFAL-TR-66-408, AF Avionics Laboratory, Wright Patterson Air Force Base, Ohio.

Peckham, J.B. (1979), Direct Voice Input (DVI) to Avionic Systems: A Proposal for Research at RAE and a Brief Summary of Current Research in NATO Countries, unpublished manuscript from Royal Aircraft Establishment, London, England.

Pierce, J.R. (1969), Whither Speech Recognition?, Journal of the Acoustical Society of America, 46, 1049-1051.

Rabiner, L.R. and M.R. Sambur (1975) An Algorithm for Determining the Endpoints of Isolated Utterances, Bell System Technical Journal, 54, 297-315.

Rabiner, L.R. and M.R. Sambur (1976) Some Preliminary Results on the Recognition of Connected Digits, IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-24, 170-182.

Reddy, D.R. (1967), Computer Recognition of Connected Speech, Journal of the Acoustical Society of America, 42, 329-43.

Robinson, A.L. (1979), More People are Talking to Computers as Speech Recognition Enters the Real World, Science, Vol. 203, 16 February, 1979, pp. 634-638.

Rosenberg, A.E. and C.E. Schmidt (1977), Recognition of Spoken Spelled Names Applied to Directory Assistance, The Journal of the Acoustical Society of America, Vol. 62, Supplement 1, S63.

Rosenberg, A.E. and C.E. Schmidt (1978), Directory Assistance by Means of Automatic Recognition of Spoken Spelled Names, Proceedings of the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK.

Sambur, M.R. and L.R. Rabiner (1975), A Speaker-Independent Digit-Recognition System, Bell System Technical Journal, 54, 81-102.

Sambur, M.R. and L.R. Rabiner (1976), A Statistical Decision Approach to the Recognition of Connected Digits, IEEE Transactions on Acoustics, Speech and Signal Processing, Volume ASSP-24, Number 6, 550-558.

Scott, P.B. (1976), Voice Input Code Identifier, Final Technology Report, Air Force Contract F30602-75-C-0111, Report Number RADC-TR-77-190, Rome Air Development Center, Air Force Systems Command, Griffiss AFB, NY.

Sondhi, M.M. and S.E. Levinson (1977), Relative Difficulty and Robustness of Speech Recognition Tasks that Use Grammatical Constraints, Journal of the Acoustical Society of America, 63, Supplement 1, S64(A).

Sondhi, M.M. and S.E. Levinson (1978), Computing Relative Redundancy to Measure Grammatical Constraint in Speech Recognition Tasks, Proceedings of the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK.

Tsuruta, S. (1978), DP-100 Voice Recognition System Achieves High Efficiency, JEE (Japanese Magazine), DEMPA Publications, Japan, July, 1978, 50-54.

Vicens, P.J. (1969), Aspects of Speech Recognition by Computer. Technical Report, Stanford University, AI Memo 85, Stanford, CA, (Ph.D. Dissertation).

Welch, J.R. (1977), Automatic Data Entry Analysis, Final Technical Report RADC TR-77-306, Rome Air Development Center, Rome, NY.

White, G.M. and R.B. Neely (1975), Speech Recognition Experiments With Linear Prediction, Bandpass Filtering, and Dynamic Programming, IEEE Transactions Acoustics, Speech, and Signal Processing, Volume ASSP-24, 183-188.